申报	系列：教师系列教学科研并重型
	专业：管理科学与工程
	职称：副教授

业绩成果材料

（申报人的业绩成果材料包括论文、科研项目、获奖以及其他成果等）

单　位（二级单位） 数学与信息学院

姓　名 古万荣

材料核对人：　　单位盖章：

核对时间：

华南农业大学制

目　录（模板）

一、教学研究业绩

1.教学研究项目：关于人工智能时代的Spark教学资源开发项目的立项通知（合同）及有关佐证材料 1
2.教学研究项目：关于大数据挖掘与可视化课程的线上线下混合教学模式应用与实践项目的立项通知（合同）及有关佐证材料 6
3.教学研究项目：关于云计算与大数据第四章（数据分析专题）校级思政教改项目立项通知 12
4.教改论文：计算机专业课互动教学方法及其效果分析.. 23
5.编写教材：《C语言程序设计入门到精通》 29
6.编写教材：《商务数据挖掘与可视化》 41

二、科研项目

1.主持：基于超图模型的新闻推荐及应用研究的立项通知及有关佐证材料 47
2.主持：基于复杂对象关系模型的新闻推荐研究及应用项目的立项通知（合同）及有关佐证材料 48
3.主持：协同超图推荐模型研究及应用项目的立项通知（合同）及有关佐证材料 51
4.主持：复杂多维数据的时空监控方法研究及其在经济社会中的应用项目的立项通知材料 62
5.主持：基于深度学习及混频数据模型的流行病组合预测方法立项通知书及有关佐证材料 64
6.主持：大数据智慧种养平台关键技术研究及其应用示范

（合同）及有关佐证材料 66
7.主持：基于深度神经网络的猪只行为识别与分析研究项目的立项通知（合同）及有关佐证材 82
8.主持：衢州柑橘大数据分析建模平台（合同）及有关佐证材料 98
9.主持：智慧两化党支部大数据平台（合同）及有关佐证材料 108
10.主持：面向海量数据的 Mooze 商城推荐系统（合同）及有关佐证材料 113
11.主持：关于招标采购信息抓取与格式化研究与应用项目的结题报告及有关佐证材料 117
12.主参：风险偏好、风险认知与农户小额信贷的“自我排斥”项目有关佐证材料 124
13.主参：交通无线磁阻传感器网络深度学习去噪方法研究项目有关佐证材料 124

三、论文、著作等

1.检索证明 126
2.以第一作者发表本专业论文情况
2.1. An Improved Muti-Task Learning Algorithm or Analyzing Cancer Survival Data 131
2.2. 面向非随机缺失数据的协同过滤评分方法 149
2.3. 基于 AdaBoost 算法的药物—靶向蛋白作用预测算法163
2.4. Efficient Multi-View Clustering via Essential Tensorized Bipartite Graph Learning 174
2.5. 基于多模型优化的超声图像肿瘤自动识别..... 188
3.以通讯作者发表本专业论文情况

3.1. 推荐系统信息跨领域的改进迁移学习算法 201
3.2. An improved recommendation based on graph convolutional network 211
3.3. 创新环境指数构建及基于自适应 ARIMA 的动态预警监控研究 237
3.4. 融合评分上下文和物品相似度的推荐算法..... 250
3.5. 基于决策树改进深度交叉网络的推荐模型..... 262
3.6. 基于单核苷酸多态性筛选的品种鉴定方法..... 274
3.7. 基于点割集图分割的矩阵变换与分解的推荐算法.. 282
3.8 居民消费价格指数、工业增加率与股票指数波动关系分析基于 GARCH—MIDAS 模型 295

四、科研成果

1.知识产权
1.1.软件著作版权证书：面向 X 光图像的肺炎识别和定位系统[简称：X 光肺炎预测及定位系统]V1.0..... 308
1.2.软件著作版权证书：果实识别系统 V1.0..... 309
1.3.软件著作版权证书：基于 faster r-cnn 的甲状腺肿瘤超声图像自动识别系统 310
1.4.软件著作版权证书：基于 Nutch 的互联网信息抓取系统[简称：互联网信息抓取系统]V1.0 311
1.5.软件著作版权证书：易书网系统 312
1.6.软件著作版权证书：“百租易”物品租赁平台软件.. 313
1.7.软件著作版权证书：智能问答系统软件 315
1.8.软件著作版权证书：个人事务管理系统 V1.0..... 317
1.9.软件著作版权证书：HTML 旅游网站软件 319
1.10.软件著作版权证书：基于 iOS 平台的图像社交软件 321

1.11.软件著作版权证书：基于词向量模型的查询词扩展搜索平台 323
1.12.专利授权证书：一种基于改进的神经网络实现食品风险评估的方法 324

五、其他业绩

1.指导学生学科竞赛
1.1.2023 年度蓝桥杯国家级比赛三等奖 1 项 325
1.2.2023 年度蓝桥杯广东省级比赛一等奖 2 项 326
1.3.2023 年度蓝桥杯广东省级比赛二等奖 1 项 328
1.4.2023 年度蓝桥杯广东省级比赛三等奖 3 项 329
1.5.2025 年度全国大学生“挑战杯”三等奖 1 项 332
2.个人荣誉
2.1.院级“年度十佳科技工作者”证书 333
2.1.校级“三下乡优秀指导老师”证书 334

教育部产学合作协同育人—教学内容和课程体系改革项目合作合同

甲方：北京红亚华宇科技有限公司　　项目编号：201702139010

乙方：华南农业大学　　签订时间：2018 年 7 月

根据《中华人民共和国教育法》、《中华人民共和国职业教育法》、《中华人民共和国合同法》及其他相关法律、法规之规定，双方本着自愿平等、合作共赢、诚实信用的原则，就“产学合作协同育人-教学内容和课程体系改革”项目建设事项，签订本协议。

一、 合作原则

教学内容和课程体系改革项目旨在通过设立课程资助，协助学校进行课程建设，提升教学质量，促进符合产业需求的人才培养工作，并依据大数据、云计算等行业技术的发展趋势，制定符合行业人才需求的核心课程。

二、 合作金额及付款方式

2.1 本合同总金额为人民币（大写）叁万元整（￥30,000.00）。

2.2 付款方式：合同签订后甲方先支付人民币（大写）壹万元整（￥10,000.00）。乙方按约完成质量合格的课程内容甲方验证合格后，甲方向乙方支付其余款项人民币贰万元整（￥20,000.00）。

2.3 所有款项均使用对公转账方式，转入乙方对公账号。乙方账号信息如下：

户　　名：华南农业大学

开 户 行：广州工行五山支行

银行账号：3602002609000310520

2.4 乙方收到甲方每笔款项时均需在 3 日内开具相应付款金额的增值税专用发票或增值税普通发票。甲方开票信息如下：

公司名称：北京红亚华宇科技有限公司

税　　号：91110108592335006L

地　　址：北京市海淀区学院南路 15 号院 7 号楼 16 层 1907

电　　话：010-82205218

开 户 行：中国建设银行股份有限公司北京硅谷支行

账　　号：11001085300059622019

三、 项目周期

3.1 合同签订后项目实施有效期为半年。

3.2 交付全部课程内容的时间：2018 年 12 月 31 日

3.3 乙方应于 2018 年 12 月 31 日前将上述完成的所有课程内容交付甲方。乙方因故不能交付，应在交付期届满前 30 日通知甲方，双方另行约定交付日期。乙方到时仍不能交付，甲方可以终止合同。

四、 乙方需完成的课程内容

4.1 需要完成的内容

一级目录	二级目录	概括
人工智能	1. 无人驾驶	每个项目都包括数据采集、知识表示、自动推理、机器学习、自然语言理解、自动程序设计、应用等。（可根据项目进行修改，但是数据采集、机器学习、识别、练习、应用必须包括在内）
	2. 图像识别（基于卷积神经网络的人脸识别 Olivetti Faces 数据集）	
	3. 语音识别（单音源语音识别，HMM-GMM，数据集 Google Code project by Hakon Sandsmark））	
	4. 文字识别（在线客服、聊天机器人）	
	5. 智能理财	
	6. 无人超市（新零售模式）	
	7. 购物推荐	

4.2 内容要求：

（1）实验环境：

自行搭建 docker 镜像，如不能使用 docker 环境，需和红亚可以进行商议后再进行确定。

（2）课件制作：

①格式上：以 markdown 形式编写。课件包括实验目的、实验原理、实验环境（包括操作系统、实验工具）、实验步骤（每一步骤都需要文字加图片的形式说明，每一个步骤都需要有一个知识点考核）、

实验考核（数量 4-6 个，考核形式可单选、多选、填空）。同时每个课件都需要标明所需学时（即普通学时按照一个小时算）。

②内容结果上：保证课件质量，同时结果要和实验目的一致，对结果进行可视化、文字分析阐述或者以其他形式展现。

③课件制作会给出样本。另外每个实验还需要制作相应指导 PPT。

（3）数据：

针对不同实验所需数据需要自行抓取，对于引用的数据源需要表明来源，红亚科技不提供任何形式的数据。另外数据格式需要说明，可以以文字加列表形式进行简单说明。

（4）交付：

后期完成之后需要给红亚科技提供 docker 镜像或者 dockerfile，交付方式待定（百度云盘等其他方式）。需要提供源代码、markdown 格式课验压缩包（图片文件夹和.md 结尾的 markdown 文本文件）、实验指导 PPT。

五、 双方的权利和义务

（一）甲方的权利与义务

5.1 甲方对乙方提供相关的工作配合。

5.2 甲方对乙方在工作中提出的问题及时进行反馈并进行相关工作的配合。

（二）乙方的权利与义务

5.3 乙方同意将其提供的课程内容的所有文件的知识产权、署名权、著作权、商标权、专利权、使用权、所有权等法律权利、人身权利、财务权利等权利永久性卖断于甲方。

5.4 乙方对其交付的课程内容，需保证内容质量。质检出现的问题时需要配合改正。

5.5 乙方未经授权不得将本合同中涉及的所有内容给任何第三方使用。

5.6 乙方需保证所提供的课程内容，没有任何版权问题。

六、 保密条款

6.1 本协议及本协议所涉及的一切文件、文本、往来函件、会议纪要等，签约双方应严格保密，任何一方除为了履行必要的法定义务外，未经对方的同意，均不得以公开、声明或以暗示的方式告知签约方之外的其他任何自然人、法人及其他组织。

6.2 因本协议的履行及签约双方的商洽接触等，任一方所掌握的对方的商业秘密、商业信息、及对方明确提出希望该方保密的非商业信息，任一方应承担保密责任，不得将上述信息透露给第三方，或未经对方许可而擅自使用，或许可第三方使用。

七、 违约责任

7.1 校企双方均应按照本协议约定履行义务，一方未履行义务或者履行义务不符合本协议约定的，视为违约，应承担违约责任，违约方应向对方赔偿全部的直接和间接损失。

八、 争议的解决

8.1 依据本协议，双方确定为项目合作关系，双方中任何一方对另一方的商业行为、法律行为及经营损失不承担连带责任。

8.2 在双方的合作办学过程中，如出现任何合作问题，双方应及时协商解决。若任何一方的行为违背国家法律法规、国家相关政策，违反商业道德和社会公共道德，故意或过失造成对对方利益损害时，另一方有权提前终止或解除本合同。

8.3 本协议之变更须经双方同意并采取书面形式备案。变更或终止本协议时须提前一个月向另一方发出书面通知。

8.4 因履行本协议而发生的争议，双方应本着互谅互让的原则协商解决，协商不成时，可向合同签订地人民法院提起诉讼。

九、 其它

9.1 在本合同基础上，双方可另行签订其它合作内容的具体协议，其核心内容不得与本协议各项约定相抵触。

9.2 本合同未尽事宜，双方另行签署书面补充协议、合作方案，均具有同等法律效力。

9.3 本合同正本壹式肆份，校企双方各执贰份。

甲方：北京红亚华宇科技有限公司

（公章）

负责人签字：

签订日期：2018年7月

乙方：华南农业大学

（公章）

负责人签字：

签订日期：

华南农业大学2020年校级教改项目立项名单

项目编号	项目名称	申报人	项目类别	专项类别
JG20001	基于大规模在线开放课程的线上“金课”建设探索--以“压花艺术”课程为例	陈国菊	重点项目	金课改革
JG20002	数学类课程线上线下混合教学的探究与实践	江雪萍	重点项目	金课改革
JG20003	基于高阶性、创新性与挑战度“三位一体”的“金课”改革研究与实践——以《中国文学批评史》为例	孙宗美	重点项目	金课改革
JG20004	基于创新能力培养的专业实践课中混合教学模式的探索与实践--以《操作系统》实践课为例	王金凤	重点项目	金课改革
JG20005	“MOOC+线下课”教学模式的探索与实践—以《生物化学》“MOOC+线上见面课”课程教学为鉴	初志战	重点项目	金课改革
JG20006	分级教学背景下大学英语B级线上线下混合式写作教学模式研究	张雅娜	重点项目	金课改革
JG20007	大思政视域下高校思政课校内实践教学基地建设的研究与实践 ----以华南农业大学为例	叶晖有	重点项目	思政课程
JG20008	课程思政教育在《禽病学》教学中的应用与实践	贾伟新	重点项目	课程思政
JG20009	中国民族音乐课程中的思政教学探索	朱虹	重点项目	课程思政
JG20010	“三全育人”背景下《无机功能材料》课程思政的建设与实践	倪春林	重点项目	课程思政
JG20011	新工科背景下工程类人才培养中课程思政育人模式研究	李高扬	重点项目	课程思政
JG20012	“三全育人”理念下“课程思政”的向度与路径研究——以3门通识教育课的教学改革为例	陈晓梅	重点项目	课程思政
JG20013	基于混合式教学的课程思政改革探索——以《当代中国社会问题研究》课程为例	张运红	重点项目	课程思政
JG20014	“立德树人”视域下中外合作办学课程思政教学模式探索与实践——以日语专业中日合作办学项目为例	赵无忌	重点项目	课程思政
JG20015	艺术类史论课程融入课程思政的探索与实践	胡辉	重点项目	课程思政
JG20016	基于“新农科”建设的农林经济管理专业人才培养模式探索与实践	陈有华	重点项目	四新改革
JG20017	新农科背景下多学科交叉融合的园林专业人才培养模式研究与实践	翁殊斐	重点项目	四新改革
JG20018	新工科背景下基于能力培养的城乡规划专业“双线混融式”实践教学改革与创新	赵建华	重点项目	四新改革
JG20019	大数据背景下基于文本挖掘的新农科人才需求分析与人才培养模式改革研究	张建桃	重点项目	四新改革
JG20020	以新工科人才培养为导向的发酵工程实践教学模式探索	徐学锋	重点项目	四新改革
JG20021	新文科理念下应用型管理人才培养模式改革与实践——以土地资源管理专业为例	李灿	重点项目	四新改革
JG20022	新农科背景下农林院校旅游管理专业建设的内涵与提升路径	林敏慧	重点项目	四新改革
JG20023	面向新文科建设的农林院校历史专业改革与发展研究	赵飞	重点项目	四新改革
JG20024	新文科建设视域下的哲学本科人才培养模式改革研究与实践	谢翾	重点项目	四新改革
JG20025	基于创新能力培养的《养兔学》教学模式改革研究与实践	孙宝丽	重点项目	
JG20026	以传承中华优秀传统文化为特色的应用型音乐人才培养模式改革研究	郑颜文	重点项目	
JG20027	风景园林专业美术类课程群模块构建与教学创新	李静	重点项目	
JG20028	社会主义核心价值观教育融入“英美文学选读”教学的研究与实践	尹晓霞	重点项目	
JG20029	基于学生工程能力培养的机械基础实验课程改革与实践	甄文斌	重点项目	
JG20030	基于校企合作的风景园林协同育人模式探索与实践	黄东	重点项目	

JG20031	普通大学高水平运动员培养模式改革与研究----以华南农业大学为例	陈华东	重点项目	
JG20032	智慧植保导向下的精准农业人才培养体系构建	徐汉虹	重点项目	
JG20033	基于教育数据挖掘的高校学习评价方法研究	倪妙珊	重点项目	
JG20034	工程基础实践课在高校“冲一流”背景下的改革与实践	徐相华	重点项目	
JG20035	农林院校创新创业型人才培养视角下知识产权教育课程体系建设研究	张日新	重点项目	
JG20036	大规模远程在线考核关键技术研究	陈湘骥	重点项目	
JG20037	国际化视野下生态环境类人才培养的研究与实践	蔡昆争	重点项目	
JG20038	高等农业院校工程类专业“普通化学”在线开放课程建设与应用	刘海峰	重点项目	
JG20039	基于移动平台的自然地理野外实习虚拟仿真系统建设与研究	钟晓兰	重点项目	
JG20040	“植物学类专业课+古诗词+立德树人”的三全育人教学模式探索	解新明	重点项目	
JG20041	植物学虚实协同融合实验教学体系创新路径与实践	方颖	重点项目	
JG20042	法庭科学技术虚拟仿真实验项目的研究与实践	杜国明	重点项目	
JG20043	基于TPC 模式的温氏班《养牛学实训》课程改革与实践	邓铭	重点项目	
JG20044	校企协同育人下的应用型知识产权人才培养创新模式研究	王红	重点项目	
JG20045	DMap在大学英语的教学改革及研究	钟志英	重点项目	
JG20046	基于创新能力培养的遗传学实验教学资源体系改革研究	李楠	重点项目	
JG20047	产教融合视域下复合应用型旅游管理人才培养体系创新与实践	陈丽丽	重点项目	
JG20048	农业高校广播电视编导专业涉农人才培养模式研究	郑文华	重点项目	
JG20049	动画专业产教融合协同育人模式的研究	陈瑰丽	重点项目	
JG20050	基于产教深度融合的农林经济管理实践教学体系内涵建设	李巧璇	重点项目	
JG20051	高等农业院校与农业产业园合作育人模式创新研究——以华南农业大学为例	潘军	重点项目	
JG20052	核心素养助推智慧教育导向的《水产动物组织胚胎学》教学探索与实践	范兰芬	重点项目	
JG20053	“后疫情时代”线上线下混合式教学模式在病毒学课程中的应用	马金成	一般项目	金课改革
JG20054	“线上线下”混合式教学中协作学习体系的构建——基于经管专业数据分析类课程的探索与实践	李宗璋	一般项目	金课改革
JG20055	《植物生理学实验》微课的建设	钱春梅	一般项目	金课改革
JG20056	后疫情时期Orff教学法在音乐通识课“线上”教学模式研究与实践	李伟	一般项目	金课改革
JG20057	基于 MOOC+SPOC 线上线下混合模式的林学实验教学改革与实践	苏艳	一般项目	金课改革
JG20058	互联网+背景下《电工电子技术》 教学中的微课运用研究与实践	史艳玲	一般项目	金课改革
JG20059	新冠疫情后基于中国大学MOOC异步SPOC的混合式教学模式推进策略与效果评价——三个“三位一体”视野下对两门公共管理类课程的对比研究	吴法	一般项目	金课改革
JG20060	基于MOOC 的SPOC混合教学模式在《动物学》教学中应用探索	佟富春	一般项目	金课改革
JG20061	线上翻转课堂的教学模式在操舞类体育课中的应用研究	何灵捷	一般项目	金课改革
JG20062	基于网络教学平台的大学英语翻译模块课程混合式教学改革研究与构建	仇如慧	一般项目	金课改革
JG20063	“林木遗传育种学”课程的“线上”+“线下”混合式教学模式研究与实践	欧阳昆唏	一般项目	金课改革

JG20064	基于“问题引导”的《大学物理实验》课程线上线下结合教学模式改革和实践研究	王先菊	一般项目	金课改革
JG20065	“线上”和“线下”有机融合的SPOC混合式教学在模拟电子技术实验中的研究与实践	成兰仙	一般项目	金课改革
JG20066	基于“学习通+翻转课堂”的混合教学模式的创新与实践--以《数据库应用》公共课为例	郭艾侠	一般项目	金课改革
JG20067	《工程测量学》混合式教学模式研究	林观土	一般项目	金课改革
JG20068	基于SPOC的翻转教学模式的研究与实践---以数字通信原理课程为例	胡春筠	一般项目	金课改革
JG20069	《测量学》课程混合式教学改革研究	刘建滨	一般项目	金课改革
JG20070	“清洁生产与循环经济”线上和线下混合式教学研究与实践	蒋成爱	一般项目	金课改革
JG20071	大数据挖掘与可视化课程的线上线下混合教学模式应用与实践	古万荣	一般项目	金课改革
JG20072	“互联网+教育”环境下控制工程基础课程混合教学改革与实践	邢航	一般项目	金课改革
JG20073	高等数学翻转课堂探究式教学方式的创新与实践	何婧	一般项目	金课改革
JG20074	高校思政“基础”课专题式教学研究	郑萍	一般项目	思政课程
JG20075	习近平相关论述指导下高校思政课在新时代人才培养中的作用研究	王景奇	一般项目	思政课程
JG20076	以"传统文化"为切入点的课程思政融入金属工艺课程教学模式探索	潘子广	一般项目	课程思政
JG20077	课程思政下的专业课程教学设计——以生态道德教育为导向的动物学课程 “线上 +线下 ”混合教学探索	曾芳	一般项目	课程思政
JG20078	“课程思政”视域下农事训练课程与思政教育协同育人路径研究	陈娴	一般项目	课程思政
JG20079	大思政背景下的英语视听说混合式课程建设	秦建华	一般项目	课程思政
JG20080	土壤类课程思政的探索与实践：多学科与多维度之融会贯通	卫泽斌	一般项目	课程思政
JG20081	新形势下“芯”专业课程思政探索与实践	代秋芳	一般项目	课程思政
JG20082	公共基础课无机&分析化学及实验课程思政元素的挖掘与实践研究	周晓华	一般项目	课程思政
JG20083	融合思政元素和翻转课堂的线性代数教学模式探究	朱艳丽	一般项目	课程思政
JG20084	“新农科”背景下创新型融媒体人才的教学探究——以《新媒体实务》课程为例	赵娜	一般项目	四新改革
JG20085	新工科理念下“食品添加剂”的“学-研-产”纵横延拓型课程设计与实践	肖杰	一般项目	四新改革
JG20086	新农科背景下的智慧教学环境的设计研究	周恩浩	一般项目	四新改革
JG20087	云教学条件下新工科专业实践课程在线教学模式探索与实践——以家具设计与工程专业为例	宋杰	一般项目	四新改革
JG20088	新工科背景下测绘工程实践教学改革	陈俊林	一般项目	四新改革
JG20089	新农科背景下《土壤肥料学》课程思政建设和实践教学形式改革研究	徐会娟	一般项目	四新改革
JG20090	国际化新形势与新林科背景下林学专业人才培养模式改革与探索	周玮	一般项目	
JG20091	计算机基础课程线上实验平台与资源建设应用研究	李玉峰	一般项目	
JG20092	高校教师应对在线教学的教育技术应用能力提升路径探索	李玉玲	一般项目	
JG20093	十年来国内高等农林教学研究的可视化分析	杨永忠	一般项目	
JG20094	互联网背景下 “产出导向法 （POA） ”在大学英语教学中的应用研究	牛莉	一般项目	
JG20095	基于供给侧结构性改革理论的资科专业《土壤学》课程教学改革研究与实践	任宗玲	一般项目	
JG20096	模拟法庭课程建设质量评价体系研究	张艳琼	一般项目	

JG20097	基于教育大数据的高校学生学业预警机制构建研究——以华南农业大学为例	曾鸣	一般项目	
JG20098	社会工作专业学生实习选择及分化逻辑研究——基于八个年级实习材料的分析	马林芳	一般项目	
JG20099	新媒体环境下形象设计的教学改革实践	郑丽娜	一般项目	
JG20100	“课程思政”视域下图书馆与思政教育协同育人模式研究	黄承红	一般项目	
JG20101	禁毒社会工作标准化课程体系研究与实践	曾永辉	一般项目	
JG20102	新时代背景下植物与中华文化的有机结合——《植物命名与中华文化》课程教学改革与实践	单体江	一般项目	
JG20103	基于奥尔堡 PBL 模式的《小动物传染病学》实践教学创新与研究	孙海亮	一般项目	
JG20104	R语言融入统计学课程教学的探索	潘逵	一般项目	
JG20105	基于“社会—生态实践”的风景园林设计基础类课程创新改革研究	汤辉	一般项目	
JG20106	基于超星平台的大学英语智慧课堂教学模式设计与应用研究	夏妙月	一般项目	
JG20107	人工智能技术与英语交互式口语教学的融合应用研究	于玮婷	一般项目	
JG20108	虚拟现实(VR)技术在《互联网+农业》实验教学改革中的探索	张雷	一般项目	
JG20109	基于新冠肺炎疫情形势下的《植物学实验》课程教学模式改革研究	张荣京	一般项目	
JG20110	以学生为中心的遗传学实验优化与实践	刘振兰	一般项目	
JG20111	可视化植物营养学实验课程建设	喻建刚	一般项目	
JG20112	基于实践和创新的地图学教学改革研究	吴小芳	一般项目	
JG20113	数据结构在线评判系统题库的建设与应用研究	司国东	一般项目	
JG20114	《农业机器人》教学模式与创新实践	胡练	一般项目	
JG20115	“课程思政”视角下的UI设计教学改革与创新	欧阳文昱	青年项目	课程思政
JG20116	“互联网+实体课堂”线上线下混合式体育教学模式研究	周文英	青年项目	课程思政
JG20117	工科短学时课程“线上”+“线下”混合式教学模式研究与实践—以《建筑给排水工程》为例	梁瑜海	青年项目	课程思政
JG20118	课程思政视角下的《生命科学与人类文明》教学改革与初探	李培	青年项目	课程思政
JG20119	《草坪营养与施肥》线上线下翻转课堂研究与实践	钟天秀	青年项目	课程思政
JG20120	基于虚拟仿真平台的“线上+线下”教学模式在《甲壳动物实验及实习》中的应用	何玉慧	青年项目	课程思政
JG20121	基于微课的翻转课堂教学模式在动物营养学中的运用初探	谭成全	青年项目	课程思政
JG20122	新农科背景下涉农法律人才培养研究	唐穗	青年项目	四新改革
JG20123	新工科背景下的《操作系统》课程改革研究	张猜	青年项目	四新改革
JG20124	新农科背景下基于学生创新能力培养的生物化学实验教学改革研究	许燕珍	青年项目	四新改革
JG20125	新工科背景下模拟电子技术实验教学内容与考核方式改革研究	黄慧娴	青年项目	四新改革
JG20126	BOPPPS 教学模式在创新创业导向型《财务管理》中的应用	龙思颖	青年项目	
JG20127	《园艺植物种子生产原理与技术》理论与实验课程内容优化改革研究	程蛟文	青年项目	
JG20128	基于病历挖掘式实验教学模式的改革与实践——以《兽医临床病理学实验课》为例	韩庆月	青年项目	
JG20129	以《畜牧微生物学实验》为例探索专业基础实验课教学模式改革	李鸿鑫	青年项目	

JG20130	农林类院校《保险精算学》课程教学方法改革与探索	杨明旭	青年项目	
JG20131	IT类专业“1站2导3平台”创新创业实践教育模式研究	刘树鑫	青年项目	
JG20132	以建立数据库为导向提升《水产饲料添加剂学》教学效果的研究	陈世俊	青年项目	
JG20133	斑马鱼突变模型在《遗传学实验》教学实践中的应用研究	谢少林	青年项目	
JG20134	《空间数据库原理与应用》在线社区式教学创新与实践	刘轶伦	青年项目	
JG20135	基于创新能力培养的实验类教学方式改革研究与实践	毛子翎	青年项目	
JG20136	生态学野外综合实习课程改革	李荣华	青年项目	
JG20137	基于微信公众号的基础化学实验教学辅助平台的构建与实践	赖婷	青年项目	
JG20138	基于行动导向型教学模式的《汽车理论》标准化教学组织模型研究	王昱	青年项目	
JG20139	CLIL视域下“农业＋日语”课程教学改革研究	郭圣琳	青年项目	
JG20140	翻转课堂在《鱼类学》分类教学中的应用	崔科	自筹项目	金课改革
JG20141	实践课线上线下混合式教学模式改革探索——能源与环境系统工程专业《环境工程实验》	简秀梅	自筹项目	金课改革
JG20142	“线上虚拟仿真+线下光路设计”的光信息数字化实验体系改革与优化	翁嘉文	自筹项目	金课改革
JG20143	基于腾讯课堂+SPOC的《汇编语言程序设计》翻转课堂教学模式研究	吴理华	自筹项目	金课改革
JG20144	信息化条件下高等数学混合式教学模式的研究与实践	王雪琴	自筹项目	金课改革
JG20145	基于认知主义学习理论的基础日语课程线上线下混合教学模式研究	高琴	自筹项目	金课改革
JG20146	基于翻转课堂的《畜牧微生物学》理论课教学改革	孙媛	自筹项目	金课改革
JG20147	基于微课的翻转课堂模式在《园艺植物栽培学》实验教学中的创新研究	柴喜荣	自筹项目	金课改革
JG20148	基于慕课混合式教学的数学类课程评价研究	杨志程	自筹项目	金课改革
JG20149	基于云课堂的《矿物资源与宝石鉴赏》混合式教学模式研究	李博	自筹项目	金课改革
JG20150	翻转课堂在《园艺产品物流与供应链管理》 线上线下混合教学中的应用	单伟	自筹项目	金课改革
JG20151	基于雨课堂的《无机化学》课程线上线下混合式教学改革与实践	何良	自筹项目	金课改革
JG20152	中国近现代史纲要现场教学研究	赵志宇	自筹项目	思政课程
JG20153	疫情防控常态化背景下的高校思政课教学模式研究与实践	袁尔纯	自筹项目	思政课程
JG20154	“课程思政”视域下高校思政课教学与创新创业教育协同育人机制研究	林晓燕	自筹项目	思政课程
JG20155	新时代高校思政课“混合式”教学模式研究与实践	陈洁	自筹项目	思政课程
JG20156	新媒体视域下高校思想政治理论课获得感研究 ——以《思想道德修养与法律基础》为例	余喜	自筹项目	思政课程
JG20157	《中国近现代史纲要》课程主题教学整体优化研究与实践	周新华	自筹项目	思政课程
JG20158	课程思政改革的中国话语建构研究	杨琳	自筹项目	思政课程
JG20159	基于乡村振兴和新农科建设背景下《农业环境与乡村旅游》课程思政的探索与实践	吕辉雄	自筹项目	课程思政
JG20160	农林院校《大学英语》课程思政实施路径研究	杜龙鼎	自筹项目	课程思政
JG20161	《大学物理实验》课程思政教学改革与实践探索	杨小红	自筹项目	课程思政
JG20162	计算机通识教育课程中“思政”教学建设与实践研究	万华	自筹项目	课程思政

JG20163	工匠精神视域下创新人才培养与课程思政协同育人模式研究——以中国传统工艺课程为例	张静	自筹项目	课程思政
JG20164	基于“学习强国+课程思政”的大学英语写作混合式教学模式研究	苏君	自筹项目	课程思政
JG20165	商务英语专业课“课程思政”教学模式研究——以《电子商务》为例	尹静媛	自筹项目	课程思政
JG20166	将“思政”之盐融入计算机算法类课程之汤的隐性教育探索和实践	郑婵	自筹项目	课程思政
JG20167	课程思政背景下遗传学教学改革研究与实践	陈志雄	自筹项目	课程思政
JG20168	以能力为导向的《环境影响评价》教学改革及课程思政探索	陈烁娜	自筹项目	课程思政
JG20169	新农科和学分制双背景下动物科学专业课程结构调整与建设	习欠云	自筹项目	四新改革
JG20170	新农科下高等农业院校农科人才培养错位与政策优化研究	刘春桃	自筹项目	四新改革
JG20171	新工科理念下房地产开发与管理专业人才培养模式改革与实践	游珍	自筹项目	四新改革
JG20172	新农科人才培养模式下大学生农业科技英语阅读能力的培养研究	叶茂连	自筹项目	四新改革
JG20173	新农科背景下的农科院校国家一流行政管理本科课程体系创新研究	朱生伟	自筹项目	四新改革
JG20174	新工科背景下农机丁颖班创新能力培养的研究	贾瑞昌	自筹项目	四新改革
JG20175	新农科背景下《林木遗传育种学》双语教学探索与实践	龙健梅	自筹项目	四新改革
JG20176	新理工科背景下基于虚拟仿真技术的《电路实验》教学改革与实践	王建华	自筹项目	
JG20177	智能化教学模式改革研究--以证据法学在线课程设计为例	赵蕾	自筹项目	
JG20178	基于信息化技术（SCAULab of Arts 微信公众号） 在英语专业文学课教学中的融合应用研究	侯金萍	自筹项目	
JG20179	语言智能背景下机器翻译与人工翻译的对比分析在非英语专业翻译教学中的作用	刘玉花	自筹项目	
JG20180	《插画设计》课程教学模式创新与跨界研究	彭艳霞	自筹项目	
JG20181	微观结构和形貌的认知教学在工程训练中的作用研究	张殿武	自筹项目	
JG20182	深化产教融合校企合作的育人模式——以市场营销专业为例	彭思喜	自筹项目	
JG20183	新工科背景下大学生专利创新能力培养的研究与实践	何效平	自筹项目	
JG20184	基于工程教育认证的测绘工程专业综合改革研究	刘惠明	自筹项目	
JG20185	基于岭南地域文化下 “中国建筑史”课程建设研究	姜磊	自筹项目	
JG20186	移动应用开发课程体系优化研究与实践	杨春	自筹项目	
JG20187	基于创新创业能力培养的《机械装备设计》课程教学改革与实践	夏红梅	自筹项目	
JG20188	基于CDIO和智慧农业导向的农业院校复合型人才培养模式探究	杨意	自筹项目	
JG20189	“互联网+”背景下校企联合工作室教学模式探究	徐宁	自筹项目	
JG20190	以学生为中心的《基础植物病理学》课程教学内容体系改革和优化研究	饶雪琴	自筹项目	
JG20191	智能化信息化背景下数学与程序设计能力培养相结合的教学改革探索与实践	廖彬	自筹项目	
JG20192	科教融合提升生态学专业理论课教学效果的实践探索	危晖	自筹项目	
JG20193	基于新媒体平台《大学生心理健康教育》课程建设	李佳媛	自筹项目	
JG20194	兽医药理学全英教学的实践探索	于洋	自筹项目	
JG20195	科教融合背景下省级科研平台承担通识教育课程的探索与实践	阙青敏	自筹项目	

备注：“项目编号”并非项目排序，是优先项目类别、专项类别依次排列后再进行编号。

华南农业大学文件

华南农教〔2025〕7号

关于公布华南农业大学2024年度课程思政示范项目立项名单的通知

各学院、部处、各单位:

根据《关于开展2024年度校级课程思政示范建设项目申报工作的通知》,学校组织开展了2024年度课程思政示范项目评选工作。经项目负责人申请、所在单位遴选推荐、学校组织专家评审、校内公示等程序，决定立项建设华南农业大学2024年度课程思政示范项目91项，包括课程思政试点学院4个、课程思政示范团队6个、课程思政示范课程14门、课程思政示范课堂25个、课程思政典型案例42个（名单详见附件），现予以公布。

本次立项的课程思政示范项目建设期至2026年12月,建设期内,示范项目原则上不允许更换负责人或变更项目团队成员。各项目负责人要严格按照2024年度申报通知要求，及时开展工作，加

快推进课程思政改革，确保高质量完成建设目标和任务。

请各学院充分认识课程思政改革的重要意义，认真贯彻《华南农业大学课程思政实施方案》，加强对教师的相关培训、指导、引领和支持，带动教师全员积极参与课程思政教学改革，持续深入抓典型、树标杆、推经验，全面提升本科人才培养质量。

附件：华南农业大学2024年度课程思政示范项目立项名单

华南农业大学

2025年1月24日

公开方式：主动公开

华南农业大学党政办公室2025年1月24日印发

附件

华南农业大学2024年度课程思政示范项目立项名单

序号	类别	项目名称	所属单位	项目负责人	团队成员（不含负责人）
kcsz2024001	试点学院	课程思政试点学院	动物科学学院	王文策	
kcsz2024002	试点学院	课程思政试点学院	艺术学院	张艳河	
kcsz2024003	试点学院	课程思政试点学院	数学与信息学院、软件学院	陈文艺	
kcsz2024004	试点学院	课程思政试点学院	公共管理学院	蔡茂华	
kcsz2024005	示范团队	《公共管理学》课程思政示范团队	公共管理学院	史传林	唐斌、姜国兵、方敏、曾小龙、吴彦
kcsz2024006	示范团队	藏粮于“技”，“数”说农业：数字化转型课程思政教学团队	数学与信息学院、软件学院	熊俊涛	黄文玲、韩方珍、古万荣、邓金、张建桃、宋歌、韦婷婷、邓成剑、黄丽清
kcsz2024007	示范团队	农业昆虫学课程群思政示范团队	植物保护学院	陆永跃	王磊、齐易香、岑伊静、黄振、程代凤、桑文、冼继东、吴建辉、潘慧鹏、何晓芳、汪荦荦、何娜芬
kcsz2024008	示范团队	大数据与机器学习类课程思政	经济管理学院	陈有华	何勤英、伍敬文、文乐、

		示范团队			李景荣
kcsz2024009	示范团队	水利工程类专业课程群课程思政示范团队	水利与土木工程学院	齐龙	韦未、王慧琳、杨海燕、卢玉华、余长洪、张巍
kcsz2024010	示范团队	光电信息类课程群思政团队	电子工程学院（人工智能学院）	林芳	刘金龙、欧阳强强、胡旭波、刘建斌、翁嘉文、徐初东、杨初平、林上港、曾应新
kcsz2024011	示范课程	水污染控制工程	资源环境学院	余光伟	梁瑜海、种云霄、陈澄宇、仲海涛、黄柱坚
kcsz2024012	示范课程	行政法与行政诉讼法	人文与法学学院	李燕	杨正喜、欧仁山、左卫霞、孙梦
kcsz2024013	示范课程	森林培育学	林学与风景园林学院	何茜	邱权、陈祖静、刘效东、潘澜、苏艳
kcsz2024014	示范课程	植物学实验	生命科学学院	梁祥修	白玫、李雁群、张荣京、何韩军
kcsz2024015	示范课程	瑜伽	体育教学研究部	何灵捷	田甜、周文英、王顺熙、姚叶戴、戴金明
kcsz2024016	示范课程	乒乓球（国球）	体育教学研究部	陈华东	钞飞侠、张波、李嘉鹏、赵东升、王一
kcsz2024017	示范课程	数据挖掘与大数据分析	经济管理学院	文乐	李琴、陈有华、何勤英、李景荣
kcsz2024018	示范课程	翻译理论与实践	外国语学院	陈喜华	王之杰、张欢、杨敏、严晓蓉，周彧丰

kcsz2024019	示范课程	咨询心理学	公共管理学院	钟向阳	何小芹、唐晓容、卓彩琴、吕玲玲、李锦顺、韩丽
kcsz2024020	示范课程	大学英语 IV（学术英语）	外国语学院	刘玉花	钟志英、柳青、李莉
kcsz2024021	示范课程	西方园林史	林学与风景园林学院	潘建非	夏宇、陈意微、李晓雪
kcsz2024022	示范课程	金融工程	经济管理学院	周丽云	陈标金、王雄志、林伟芬、陈晓洁
kcsz2024023	示范课程	民事诉讼法	人文与法学学院	刘万洪	张艳琼、王琳、杜国明
kcsz2024024	示范课程	农业技术经济学	经济管理学院	蔡键	薛春玲、周伟、苏柳方、易智敏，蔡轶
kcsz2024025	示范课堂	《物理化学》第四章第 4 节（单组元系统两相平衡）	材料与能源学院	刘维	陈明洁、胡航
kcsz2024026	示范课堂	《生态工程学》第五章第 1 节（湿地保护与生态修复）	资源环境学院	田纪辉	蔡昆争
kcsz2024027	示范课堂	《食品分析》第八章第 2 节（蛋白质的测定）	食品学院	徐振林	温棚、罗林
kcsz2024028	示范课堂	《生态农业工程》第七章 第 1 节（碳中和与农业绿色低碳）	资源环境学院	秦俊豪	赵本良
kcsz2024029	示范课堂	《篮球》--传接球与挡拆技术专题	体育教学研究部	郭城	张波、于江杨、黄燕玲
kcsz2024030	示范课堂	《大学生心理健康教育》第七	经济管理学院	何凯	刘思蓓、崔翱鸽

		章第 3 节（情绪与情感调节的方法）			
kcsz2024031	示范课堂	《金工实习》--电火花线切割加工技术专题	基础实验与实践训练中心	徐相华	陈海波、温威
kcsz2024032	示范课堂	《环境毒理学》第五章第 1 节（试验基础）	资源环境学院	银仁莉	
kcsz2024033	示范课堂	《空间分析原理与应用》第四章第 4 节(全局空间自相关分析)	资源环境学院	陈永康	钟晓兰、赵寒冰
kcsz2024034	示范课堂	《金工实习》--注射成型及零件检测专题	基础实验与实践训练中心	张殿武	温威、陈海波
kcsz2024035	示范课堂	《水球》--踩水与水球游戏专题	体育教学研究部	张俊龙	李梅、卢三妹
kcsz2024036	示范课堂	《大学英语III》 unit 5 Gender Equality	外国语学院	陈国华	苏君、李志英
kcsz2024037	示范课堂	《国际政治学》第二章 第 2 节（世界格局）	公共管理学院	欧阳晓东	黄剑飞
kcsz2024038	示范课堂	《C 语言程序设计》第五章--循环结构程序设计之 while 语句专题	数学与信息学院、软件学院	郭艾侠	王栋、邢仲璟
kcsz2024039	示范课堂	《乒乓球》--正手攻球技术专题	体育教学研究部	黄燕玲	吕立、张晓萍、郭城
kcsz2024040	示范课堂	《移动应用开发》第九章--丰富你的程序，运用程序多媒体	数学与信息学院、软件学院	杨春	

		专题			
kcsz2024041	示范课堂	《程序设计与算法语言》第三章第 1 部分--数据表示与数据类型专题	数学与信息学院、软件学院	廖彬	蔡贤资、张连宽
kcsz2024042	示范课堂	《社会保障基金管理》第四章（社会保障基金管理的对象专题）	公共管理学院	杨亚丽	
kcsz2024043	示范课堂	《流体力学》第七章第 2 节（黏性流体的两种流态）	水利与土木工程学院	黄俐	韦未
kcsz2024044	示范课堂	《种子的力量》第十章 第 2 节（植物生理学课堂）	生命科学学院	罗娜	刘亚林、刘慧丽
kcsz2024045	示范课堂	《英语国家概况》下册第四单元（美国政治制度专题）	外国语学院	张雅娜	侯金萍，张国俊
kcsz2024046	示范课堂	《森林培育学》第六章第 5 节（无性繁殖苗培育-扦插育苗）	林学与风景园林学院	陈祖静	邱权
kcsz2024047	示范课堂	《云计算与大数据》第 4 章（数据分析专题）	数学与信息学院、软件学院	古万荣	
kcsz2024048	示范课堂	《MEMS 及其应用》第四章 第 3 节（牺牲层技术）	电子工程学院（人工智能学院）	梁亨茂	
kcsz2024049	示范课堂	《人才测评、培训与开发》第十讲（雇佣与选拔中的测评 2 专题）	经济管理学院	陈灿	

kcsz2024050	典型案例	《新文科创新创业思维与就业力提升》第三讲 “谈谈读书：为什么和怎么做”	公共管理学院	周毅	申佐佐
kcsz2024051	典型案例	《审计学》第二章 第3节（职业道德概念框架）	经济管理学院	易智敏	
kcsz2024052	典型案例	《科技前沿与行业发展（植保）》微生物的语言系统之群体感应-Quorum Sensing	植物保护学院	崔紫宁	
kcsz2024053	典型案例	《大学生创新创业基础》--法学创新创业与国家发展、个人成长的关系	人文与法学学院	李玮舜	王琳
kcsz2024054	典型案例	《比较思想政治教育》--“颜色革命”与意识形态风险防范	马克思主义学院	韩谦	
kcsz2024055	典型案例	《生物技术与人类》--CRISPR基因编辑技术	林学与风景园林学院	张俊杰	周玮
kcsz2024056	典型案例	《Java与面向对象程序设计》第四章 第2节（类的分析）	数学与信息学院、软件学院	梁茹冰	黄小虎
kcsz2024057	典型案例	《跨文化交际》--“空间语言”	外国语学院	钟建玲	赵勇
kcsz2024058	典型案例	《兽医临床诊断学》兽医临床基本检查法之问诊	兽医学院	廖建昭	
kcsz2024059	典型案例	《机械设计基础》第一章 第3节（平面机构的自由度）	工程学院	卿艳梅	

kcsz2024060	典型案例	《大学英语 I》Unit 6 Text A Door closer, are you?	外国语学院	黄净	王莹
kcsz2024061	典型案例	《基础化学实验》--茶叶中茶多酚的提取与测定	基础实验与实践训练中心	王鹏程	郑明轩
kcsz2024062	典型案例	《算法分析与设计》--动态规划中的最长公共子序列	数学与信息学院、软件学院	张慧玲	司国东
kcsz2024063	典型案例	《遗传学实验课》--植物减数分裂过程染色体制片及观察	基础实验与实践训练中心	李亚娟	郭海滨
kcsz2024064	典型案例	《文学概论》--文学的超越性	人文与法学学院	姚海燕	王瑛
kcsz2024065	典型案例	《汇编语言程序设计》--循环程序设计	数学与信息学院、软件学院	吴理华	
kcsz2024066	典型案例	《生物多样性》--大陆漂移和动物地理分布	植物保护学院	刘卫欣	
kcsz2024067	典型案例	《农业英语阅读课程》--生物质能概况	外国语学院	郭敏	袁玮
kcsz2024068	典型案例	《瑜伽》--肩颈主题练习	体育教学研究部	田甜	何灵捷
kcsz2024069	典型案例	《Linux 系统及程序设计》Linux 发展史及国产化	数学与信息学院、软件学院	钟德祥	李舜鹏
kcsz2024070	典型案例	《复变函数与积分变换》--柯西积分公式	数学与信息学院、软件学院	刘曼莉	
kcsz2024071	典型案例	《生活营造》—中国茶的起源	艺术学院	王小玉	

kcsz2024072	典型案例	《工程热力学与传热学》第三章第 4 节（孤立系统熵增原理）	工程学院	李成杰	
kcsz2024073	典型案例	《数据挖掘与大数据分析》第三章第 2 节 （Seaborn 绘图基础）	经济管理学院	伍敬文	
kcsz2024074	典型案例	《高等数学》第五章第 1 节（定积分的概念）	数学与信息学院、软件学院	郑佳悦	
kcsz2024075	典型案例	《计算机科学绪论》第十三章信息保障与安全	数学与信息学院、软件学院	罗志坚	黄小虎
kcsz2024076	典型案例	《Linux 系统及程序设计》--Linux 程序设计原理	数学与信息学院、软件学院	黄立峰	罗浩宇
kcsz2024077	典型案例	《软件工程导论》第七章 第 7 节（黑盒测试之等价类划分与边界值分析）	数学与信息学院、软件学院	聂笃宪	刘鹏飞
kcsz2024078	典型案例	《美国简史》--美国进步主义运动	人文与法学学院	王跻崭	
kcsz2024079	典型案例	《通识管理训练--项目管理模块》--项目管理的涵义	基础实验与实践训练中心	谢虎	陈建军
kcsz2024080	典型案例	《PLC 原理与应用》--梯形图基本电路之兼顾公平与效率	工程学院	张闻宇	
kcsz2024081	典型案例	《数学分析》第九章第 1 节（定积分的概念）	数学与信息学院、软件学院	雷春林	王雪琴
kcsz2024082	典型案例	《数学建模》第三章--以微分方程、差分方程建模为例	数学与信息学院 软件学院	方平	张昕

kcsz2024083	典型案例	《物流与供应链管理》第八章第 1 节（供应链失调与牛鞭效应）	数学与信息学院、软件学院	陈炜颖	刘伟章、毛小娟
kcsz2024084	典型案例	《大学英语 I》第二册 Unit 2 Humanities: out of date?	外国语学院	荣莉	苏君
kcsz2024085	典型案例	《大学生心理健康教育》--与压力共处	植物保护学院	曹珂	梁丽梅
kcsz2024086	典型案例	《大学数学》第六章第 5 节（二重积分的概念）	数学与信息学院、软件学院	姚焕城	
kcsz2024087	典型案例	《数据库应用》第七章第 3 节（SQL 分组统计查询）	数学与信息学院、软件学院	张春玲	涂淑琴
kcsz2024088	典型案例	《兽医传染病学》第三章第 1 节（流行性感冒中禽流感防控）	兽医学院	代曼曼	
kcsz2024089	典型案例	《流体力学及其工程应用》第二章第 2 节（流体静平衡微分方程及其积分）	工程学院	罗远强	钟南
kcsz2024090	典型案例	《水产品加工技术》--水产加工原材料的特点	海洋学院	杨敏	李雪竹
kcsz2024091	典型案例	《食品机械与设备》--碟片式离心分离机械	食品学院	关甜	司徒文贝、宋贤良

ISSN 1007-1423
CN 44-1415/TP

现代计算机

http://www.moderncomputer.cn 欢迎浏览

中国期刊数据库CNKI全文收录期刊
中国学术期刊（光盘版）收录期刊
中文科技期刊数据库全文收录期刊
中国核心期刊（遴选）数据库收录期刊
中国学术期刊综合评价数据库收录期刊

2019 7
7月15日出版 月



目　录

研究与开发

03 基于 BP 神经网络的汽车售后业务成本预测 …… 褚福舜,吴奇石,王浩雨
10 基于 Naive Bayes 的 P2P 平台评论研究 …… 曾政多
14 光电复合高压电缆载流量计算方法研究综述 …… 谢治
18 一种混合重心重构花授粉改进算法 …… 陈昌兴,王建彬,陈建平

图形图像

23 基于残差网络的学生课堂行为识别 …… 蒋沁沂,张译文,谭思琪,杨耀祖
28 目标检测算法在乳腺癌病灶的影像学诊断上的应用 …… 陈志刚,黄斯彤
32 针对真彩图的索引色提取技术的研究 …… 张宏桥,曾晓华,郭旭升,李翔
38 基于中值滤波的图像边缘去噪算法研究 …… 郝波,卢小平
42 基于数字水印的医学图像信息隐藏系统 …… 刘泳彬,徐涛,苏锐豪,陈炯缙

教学园地

46 高职后期剪辑课程线上线下混合式教学模式设计与实践 …… 赵学华,孙慧,王丹丹,张谦,于成龙,詹峰
50 基于 GOTI 的数字媒体专业课程育人模式研究 …… 张洪民,张春迎,张英
55 计算机专业课互动教学方法及其效果分析 …… 古万荣,肖磊,谢贤芬,毛宜军,梁早清
59 面向计算思维的《Java 程序设计》课程教学改革研究 …… 高贤强

实践与经验

63 汽车 B 柱下端安全带卷收器安装支架强化措施 …… 解跃青,陈瑜
66 信息系统测评与信息系统审计 …… 张子良

开发案例

70 基于 4G 的车联网信息采集终端的设计与实现 …… 张宇,赵利,蔡成林,唐俏笑
73 基于数据挖掘和分析的食品安全智能测评系统 …… 符雨童,聂笑一,肖毅
78 基于 Wi-Fi 探针的客流统计分析的研究 …… 张国林,罗回彬,郑泽江,董晋源,刘永怡
82 基于 MQTT 协议的扫码支付中间件设计与实现 …… 崔阳
86 智慧校园人脸识别系统的设计与实现 …… 邱城铨,罗回彬,陈汕,李宇栋,苏敬雄,吕翔
91 一种 P2P 分布式文件共享方式的实现与优化 …… 刘睿,任晓慧,于波
96 基于区块链技术的高校企业实践教学管理系统 …… 肖建宏

期刊基本参数:CN44-1415/TP*1984*M*16*100*zh*P*¥30.00*1200*22*2019-07

文章编号:1007-1423(2019)20-0055-05 DOI:10.3969/j.issn.1007-1423.2019.20.012

计算机专业课互动教学方法及其效果分析

古万荣[1],肖磊[1],谢贤芬[2],毛宜军[1],梁早清[1]

(1.华南农业大学计算机工程系,广州 510642;2.暨南大学统计系,广州 510632)

摘要:

随着人工智能时代的到来,计算机的教学逐步转化为基于数学模型、面向海量数据和高性能计算的学科。但由于大多数计算机专业教学单一,导致这种转换很难随着工业环境的变化而变化。教学方法是在社会发展的大背景下不断探索磨合发展的,尤其是在技术更新极快的计算机专业教学上。通过分析国内高校的计算机教学及其存在的问题,充分考虑到计算机专业技术更新换代快的特点,提出以互动教学为主的教学培养模式,并对两年来的教学效果进行分析比较,以期培养出适合当前时代环境的计算机专业人才,并为该专业的教学改革和实践提供思路和参考。

关键词:

互动;计算机教学;教学改革;新工科

基金项目:

广东省高等教育质量工程项目(粤高教函[2018]1号)

0 引言

计算机行业的发展迅速众所周知,高等教育中的计算机专业教学常常面临着这样的囧境:大一给学生讲授的最新技术,等学生毕业后发现该技术已经较为陈旧,甚至已经被淘汰了。总体来说,目前我国高等学校包括计算机及其相关专业在内的工科专业的教学方法还存在着如下的问题有待于进一步改善。主要包括:①教学方法和内容单一陈旧,“讲风太盛”;②忽视学生分析问题和解决问题能力的培养;③单纯讲抽象理论,不重视实践性教学环节;④对学生的科研活动重视不够;⑤成绩评定不科学。

因此,计算机学科的教学及其他学科的教学早就提出了互动教学的概念,但真正实行的不多,仍然以理论讲解和布置习题实验为主。即便有一些互动教学的尝试,也没有及时对互动教学进行深入效果的对比分析。

针对以上问题,本文对计算机及其相关专业的教学方法进行全方位研究,设计面向计算机专业合理的互动实践教学方法,并通过两年来的数据积累,进行出勤率、满意度等多方面的数据对比分析。

1 国内外相关研究

计算机专业教学面临着人工智能时代社会发展带来的挑战和机遇,国内外的教学研究总体如下:

(1)国外研究现状

在美国高校教学中,计算机专业教师很多是从一流企业而来,具备博士学位,具有很强的实践能力,也能很容易地返流到企业。他们更多地注重计算机技术的实用性、先进性,鼓励学生突破技术门槛,和学生能平等交流技术和工程问题。

英国高校的计算机教学注重引导学生多思考、多参与,注重学生实践参与在教学中的比重。

德国则比较重视前沿技术的应用问题,教师入职时,有相关的培训和考核。其教学方法的主要特点是:①重视讨论法“Seminar”的运用;②重视演示法和实验法的运用;③重视学生自学。

日本高校计算机专业主要教学特点是:①讲授与

实践相结合;②注重个别化教学方法;③注重学生自学方法。

(2)国内研究现状

我国高校教学方法的结构是以传统教学方法为主。即选定一本教材、制作一份课件(可能多年不换)、以教师主讲为主。虽然随着高校教学改革的发展,在制度上引入了评教机制、在课程设计上将实验教学作为单独学分设立,但由于各方面原因,评教使得中国教师不敢对学生过于严厉,实验教学一刀切导致有些课程实验课流于形式。

近年来,我国高校计算机专业在一些学校的争取下,已经逐步突破,如注重新技术,注重吸收企业高学历实践型人才。但对于大多数院校而言,仍然面临着重重困难。

因此,鉴于国情和文化背景不同,我们高校的计算机专业教学改革,不能生搬硬套国外成功模式,需要以教学点(学校或院系)为单位,进行互动教学改革,以期达到最优化目的。

2 互动实践教学理念的整体思路

2.1 定位与构想

兴趣是学习之母,我们教学的目的是指激发兴趣(乐),以学习前沿计算机技术为目的(学),以培养会思考、会自主学习新技术的人才为落脚点(思),为国家培养更多创造型人才(创)。

2.2 互动教学方法在计算机专业课程中的设计与实践

计算机专业教学和其他专业有一定的共通之处,但由于计算机专业知识更新换代极快,新工具和新技术日新月异,工业界和学术界协同性很强。因此,互动教学方法在计算机专业课程中的设计和实现在注重教学原则性问题的同时,也要因地制宜,适当根据专业特点进行调整。

(1)激发兴趣要贯穿教学全程

传统的高校教学往往会有这样的现象,前三周逃课人数极少,但随着时间推移,逃课人数渐增,到期末最后一两周人数又多了起来。这是因为前三周学生们兴趣较高,兴趣随着时间的推移而下降。期末考试前一两周人数突增并非兴趣来了,而是基于功利目的,如考试划重点等。激发兴趣的关键点在于,让学生知道该知识点的应用落脚点,理论联系实际。有条件的教师还可以联系相关企业进行展示教学。

(2)课程内容要接轨前沿技术

许多专业知识与前沿技术脱轨,导致企业需求和高校培养对接不上,这也是导致学生厌学的一个重要因素。院校除了聘请一定配比的企业高学历人才充实教师队伍外,还可以对传统教师进行前沿技术培训,确保他们的专业知识保持新鲜度,而不是单考察论文数量或依赖于评教结果。

(3)培养会思考会学习的计算机专业人才

计算机知识的快速更新,除了我们要做好知识更新外,还需要培养具备自主学习能力的人才,确保他们毕业后有能力持续学习。主要通过以下方式:

①问题式教学:以技术疑问的方式给学生带入技术专题,引导学生发现问题或提出问题。同时,也要注意完成提出问题、分析问题、解决问题并提出更多改进意见的迭代闭环。

②启发式教学:启发式教学虽然没有明确的提出问题,但也是属于引导学生的一种方式,可以通过不同的技术方案、解决算法等给学生留下对比、分析和思考的空间。

③互动式教学:互动式教学是传统教学方法较缺乏的,包含但不仅限于教师和学生之间,学生与学生之间,还可以包含企业人员和学生之间,教师与教师之间。互动式教学有利于提高学生的自主性和参与感。

④小组合作式教学:通过划分小组,可以让较好的学生将学习经验传授给较差的学生,这种传授比单纯教师传授给学生效果要好。但该教学方式中,需要注意防止学生滥竽充数,确保每个学生都有合作参与,具有自己的贡献和对整体小组的成果有熟悉的掌握。

(4)创造是计算机专业的必备能力

在教学中,注重引导学生的创造力。计算机专业不管是软件还是硬件,具备创造力的人才才是行业的推动力。在这个环节中,需要注意:①引入开放型课题,培养创新思维;②积极开展探究式教学法。

3 传统教学与互动实践教学的效果对比

笔者及其团队挑选了三门课程作为教学对比的对象:《C 语言程序设计》、《软件开发基础(C++)》以及《云计算与大数据》。这三门课程分别是计算机专业的公

共基础课、专业基础课和专业选修课，具有代表性。

传统教学方法：理论讲解+实验课。即 32 学时用于讲解书本理论，主要是概念讲解、算法和举例；16 学时用于实验教学，主要是完成课后习题上机操作，形成代码和实验报告。

互动教学方法：应用和理论讲解共 32 学时，每次理论课之前有应用举例，并关联企业界的应用场景和可扩展性讲解，聘请企业技术人员（5 年以上一线技术研发经验）介绍与课程相关的企业应用，占 2 学时。实验课改为以小组合作式教学实验，8 个学时完成课后实验习题，8 个学时设置开放式实验课题，供学生自选。

学生选课人数的数据如表 1 所示。

表 1 学生选课人数统计表

选课人数	C 语言程序设计	软件开发基础（C++）	云计算与大数据
2017 学年（传统方法）	187	398	89
2018 学年（互动教学）	165	452	86

表 1 中的数据是选课系统中导出的数据，在评教调查中，评教回收人数和选课人数不同，则默认剔除无效评教，在部分评教项中，如果没有作答，也不计入实验统计数据中。

3.1 出勤率

为了排除人工点名或签到的方式导致学生强制性上课，影响实验数据的真实性，本实验中的人数统计都是通过照片拍摄的事后统计得出。所有课程在第一周时，跟学生约定，出勤情况不计入期终考试总评成绩。

从图 1 中可知，有以下现象和分析：①互动教学的出勤率明显较高，特别是在教学中期（第 5 周和第 10 周），差距特别明显，这说明互动性和趣味性会提高学生主动参与度。②不管是互动教学还是传统教学，在教学中期出勤率都会下降，这是由人的学习心理导致的。在学习的中期更容易产生疲倦感。③跟学生约定了出勤不计入考核分数，则出勤率普遍都不高，这也说明强制性是出勤率的一个有效保证。④有些课程出勤率在期末会反弹比较多，如软件开发基础（C++）课程、云计算和大数据，该课程难度较大，学生更希望能在考试前获得重点知识范围。

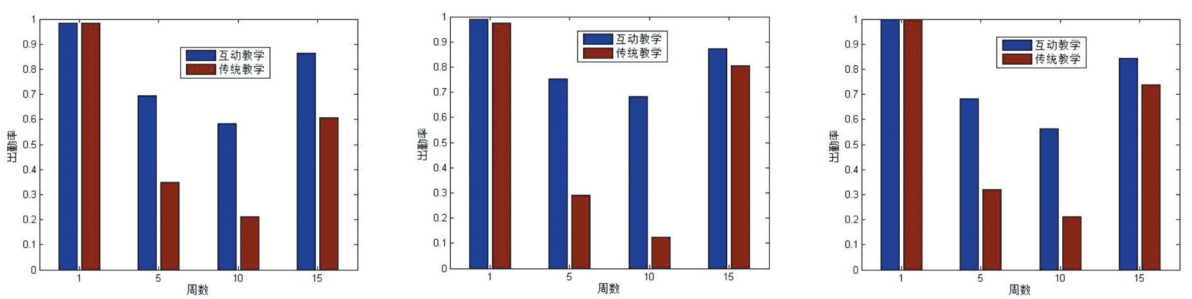
3.2 评教反馈

我单位教务处有评教系统，要求学生要进行评教后才能查询期末考试成绩，这就确保了学生不会因为自己成绩的好坏来评价教学过程。评教采用的是盲评方式，即教学教师不知道每个学生的具体评教给分，只知道统计结果。评教是使用选项式系统填空，分为 5 个等级，并对应生成相应的分数：优-95 分，良-85 分，中-75 分，合格-65 分，不合格-55 分。获得分数后，系统会自动去除最高分的前 5%样本以及最低分的后 5%样本，然后再统计出平均分作为满意度。

表 2 评教结果反馈

综合满意度	C 语言程序设计		软件开发基础（C++）		云计算与大数据	
	满意度	排名	满意度	排名	满意度	排名
2017 学年（传统方法）	90.14	79.58%	89.21	77.21%	90.24	61.65%
2018 学年（互动教学）	92.14	45.21%	93.51	32.14%	93.84	24.65%

表 2 的满意度即学生评教反馈表中的平均得分，排名是该课程在学院的排名率，如 20%，则表示在 100 个教学教师中，排名 20 左右。从表 2 的评教结果反馈统计表中可知：①使用互动方法的满意度有了明显的提高，尤其是软件开发基础（C++）和云计算与大数据这



（a）C 语言程序设计　（b）软件开发基础（C++）　（c）云计算与大数据

图 1 出勤率实验结果统计图

两门应用技术性较强的科目，这也说明了互动教学方法更受学生欢迎。②评教的分数普遍起点较高，这是因为学生都会不自觉地给自己教师一个比较高的评分起点，但评分的提升也是有参考意义的。③云计算与大数据课程人数较少，更容易在评教满意度上获得高分，并排名较前。

4 激励和考核评价机制

（1）构建与互动教学法相适应的激励机制，提高教学教师的地位和相关经费支持。

目前，高等教育严重倾斜于科研，对教学过于轻视，而大学本科生的培养质量和教学是强相关的。资源的倾斜导致许多教师不愿意过多时间花在教学上，也有些教师为了完成科研指标，聘用本科生搞科研而忽视了本科生的基础知识的重要性。

（2）从教师评估和学生自我评估两个方面，构建与互动教学法相适应的考核评价机制。

教师的教学质量评估，不能简单依赖于评教机制，其副作用在于对学生宽松的教师，往往会得到较高评价，这样导致相关教师对学生过于放松。也不能过于依赖教改论文篇数，毕竟理论和教学实践差距是非常大的。需要一个更互动的考核机制，如同行评价，互联网公开授课评价等等多元手段相结合的评价机制。

5 结语

本文探讨互动实践教学法国内外研究的基础现状，给出了互动教学的基本思路框架以及在计算机专业教学中的扩展。目前，教学研究多面向于多渠道教学（课堂教学+互联网教学），本文引导和倡议回归互动教学，即不仅仅借助于多渠道教学，更侧重于学生对教学内容本身的兴趣度、交互性、趣味性、实用性等互动协同开展。在互动设计合理的情况下，再通过多渠道工具进行扩展，以达到最优化教学效果。

参考文献：

[1]刘翠翠. 基于就业导向的计算机教学实践改革探究[J]. 科技展望，2017，27(10).

[2]马选宝. 项目教学法在计算机教学实践活动中的应用研究[J]. 无线互联科技，2017(17):77-78.

[3]赵鹏，韩素青. 基于高校计算机教学实践与教学改革的理论分析研究[J]. 教育现代化，2017(38):94-95.

[4]黄嵩. 基于任务驱动法的计算机教学实践探究[J]. 数字通信世界，2018，162(06):203.

[5]李希敏. 构建面向岗位的计算机专业教学模式——评《计算机教育教学课程研究与实践》[J]. 中国教育学刊，2018，303(7):136.

[6]兰春霞，任君庆. 游戏化学习理念在翻转课堂教学模式中的应用——以高职计算机基础课程为例[J]. 职教论坛，2017(14):56-60.

[7]朱艳辉，童启，刘强. 基于创新训练平台的计算机类专业实践教学体系研究[J]. 中国大学教学，2018，332(4):84-87.

[8]王建明，陈仕品. 基于线上课程和工作室制度的混合式教学实践研究[J]. 中国电化教育，2018(3).

[9]温武，李鹏，郭四稳，等. 基于人才联盟的"实验室+企业"网络工程实践教学基地构建[J]. 高等工程教育研究，2017(1):55-60.

[10]牟凤瑞. 基于大数据的计算机信息处理技术应用与实践[J]. 中国教育学刊，2017(S1):19-21.

[11]谭志虎，秦磊华，胡迪青. 面向系统能力培养的计算机专业实践教学模式[J]. 中国大学教学，2017(9).

作者简介：

古万荣（1982-），男，广东广州人，博士，讲师，研究方向为互联网大数据处理与分析、信息推荐模型

通信作者：肖磊（1972-），男，山东泰安人，博士，副教授，研究方向为大规模信息化架构、大数据处理与建模，Email: lein_xiao@scau.edu.cn

谢贤芬（1983-），女，广东韶关人，博士，副教授，研究方向为宏观经济统计分析、应急管理

梁早清（1970-），男，湖北仙桃人，博士，讲师，研究方向为农业信息化

毛宜军（1979-），男，广东广州人，博士，讲师，研究方向为生物基因大数据

收稿日期：2019-05-14　　修稿日期：2019-05-22

（下转第 62 页）

上册 开发入门篇



C语言程序设计
从入门到精通

何亦琛/古万荣 著

从编程菜鸟，到C语言大神的实用宝典！

- 资历深：凝聚笔者10多年的开发经验
- 内容全：囊括了C语言的200多个知识点
- 实例多：书中包括187个实例程序
- 上手快：适合初学者的详细步骤、详细代码、详细解释



101道求职中的常见智力题
117道C/C++面试题
236个C语言典型实例
337个C语言函数功能速查
Linux下C语言编程大全
SQL结构化查询语言使用大全



C语言二级考试教学视频，420分钟
C语言零基础入门视频，640分钟
C语言专题精讲，360分钟
C语言基础知识细化视频，70分钟

中国工信出版集团
电子工业出版社

内 容 简 介

本书以C语言从基础到核心开发所涉及的技术为主线，通过手把手案例指导的方式，让零基础的C语言开发人员也能轻松入手。

本书共4篇19章，分上、下两册，涵盖的主要内容有：C语言的基础、C语言开发环境、流程控制、使用数组、使用函数、控制指针、链表的概念与基本操作、文件操作、常用算法与函数、C语言预处理机制、C语言网络编程及几个综合实例。

本书内容翔实、行文活泼、实例经典、实用性强，特别适合各个层次想要深入学习C语言开发技术的人员阅读，尤其适合那些零基础而又想要快速成为C语言程序员的读者参考阅读。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目（CIP）数据

C语言程序设计从入门到精通 / 何亦琛，古万荣著. —北京：电子工业出版社，2018.1

ISBN 978-7-121-32705-6

Ⅰ. ①C… Ⅱ. ①何… ②古… Ⅲ. ①C语言－程序设计 Ⅳ. ①TP312.8

中国版本图书馆CIP数据核字（2017）第226585号

策划编辑：徐　岩
责任编辑：张　毅
特约编辑：赵树刚
印　　刷：三河市鑫金马印装有限公司
装　　订：三河市鑫金马印装有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路173信箱　　邮编：100036
开　　本：787×1092　1/16　印张：37.75　字数：966千字
版　　次：2018年1月第1版
印　　次：2018年1月第1次印刷
定　　价：88.00元（共2册）

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至zlts@phei.com.cn，盗版侵权举报请发邮件到dbqq@phei.com.cn。

本书咨询联系方式：（010）57565805。

目 录

第 1 篇 入门篇

第 1 章 从零开始认识 C 语言 1

1.1 认识 C 语言 1

1.1.1 C 语言的发展过程 1

1.1.2 C 语言的标准 2

1.2 C 语言在不同领域的扩展应用 3

1.2.1 单片机方面 4

1.2.2 LabVIEW 方面 4

1.2.3 LabWindows/CVI 方面 6

1.2.4 MATLAB 方面 7

1.2.5 JavaScript 方面 7

1.2.6 并行计算方面 8

1.2.7 Linux 平台 8

1.3 搭建 C 语言开发环境 8

1.3.1 Turbo C 9

1.3.2 Visual C++ 9

1.3.3 Linux 下的 GCC 12

1.4 第一个 C 程序 13

1.4.1 创建项目 14

1.4.2 输入源代码 14

1.4.3 编译链接程序 15

1.4.4 调试程序 16

1.5 C 程序构成分析 16

1.5.1 头文件 17

1.5.2 主函数 17

1.5.3 其他组成部分 18

1.6 拓展训练.....22
1.6.1 训练一：打印字母 C.....22
1.6.2 训练二：scanf()函数的应用.....22
1.7 技术解惑.....23
1.7.1 C 语言的一些基础规范.....23
1.7.2 C 语言的注释形式.....25
第 2 章 数据是一切程序存在的基础.....26
2.1 C 语言中的数据类型.....26
2.1.1 基本类型.....27
2.1.2 构造类型.....27
2.1.3 指针类型.....28
2.1.4 空类型.....28
2.2 数据的存储原理.....28
2.2.1 内存单元.....28
2.2.2 字符的存储.....29
2.2.3 数值的存储.....30
2.3 标识符和关键字.....30
2.3.1 标识符.....30
2.3.2 关键字.....31
2.4 常量.....32
2.4.1 直接常量.....32
2.4.2 符号常量.....32
2.5 变量.....34
2.5.1 变量声明.....34
2.5.2 变量初始化.....36
2.6 整数类型.....37
2.6.1 整数类型及其存储.....37
2.6.2 整型常量的表示.....40
2.6.3 整型数据输出.....42
2.7 字符类型.....44
2.7.1 字符常量.....45
2.7.2 字符变量及其初始化.....45
2.7.3 转义字符.....46
2.7.4 字符型数据的输出.....47
2.8 实数类型.....48
2.8.1 实数类型及其存储.....49

2.8.2 实型常量的表示.....50
2.8.3 实型变量.....51
2.8.4 实型数据的输出.....53
2.9 混合运算及类型转换.....54
2.9.1 混合运算.....54
2.9.2 强制数据类型转换.....55
2.10 逻辑型（_Bool）.....56
2.11 复数类型（_Complex 和_Imaginary）.....56
2.12 拓展训练.....57
2.12.1 训练一：变量初始化.....57
2.12.2 训练二：不同数据类型转换为十进制.....58
2.13 技术解惑.....59
2.13.1 整型数据机内存储形式.....59
2.13.2 字符型数据机内存储形式.....59
第 3 章 C 语言代码的基本组成.....60
3.1 表达式.....60
3.1.1 简单表达式.....60
3.1.2 逗号表达式.....61
3.2 运算符.....62
3.2.1 运算符概述.....62
3.2.2 算术运算符.....62
3.2.3 赋值运算符.....68
3.2.4 关系运算符.....70
3.2.5 逻辑运算符.....71
3.2.6 位运算符.....73
3.2.7 条件运算符.....73
3.2.8 其他运算符.....75
3.3 表达式的运算顺序.....76
3.3.1 运算符优先级.....76
3.3.2 运算符的结合性.....77
3.3.3 自增、自减运算符注意事项.....78
3.4 语句.....79
3.4.1 语句书写方式.....79
3.4.2 表达式语句.....81
3.4.3 空语句.....81
3.4.4 复合语句.....82

3.4.5 标号语句 82
3.5 拓展训练 83
3.5.1 训练一：加 1 和减 1 运算符 83
3.5.2 训练二：三目运算符 84
3.5.3 训练三：逗号运算符 85
3.6 技术解惑 86
3.6.1 赋值及类型转换 86
3.6.2 sizeof 是一个运算符吗 87
第 4 章 控制输入/输出的形式 88
4.1 格式化输出——printf()函数 88
4.1.1 printf()函数的格式 88
4.1.2 printf()函数的格式字符 90
4.1.3 修饰符 91
4.1.4 printf()函数实例 94
4.1.5 动态设置输出宽度和精度 99
4.1.6 printf()函数的返回值 100
4.1.7 理解输出列表 101
4.2 格式化输入——scanf()函数 104
4.2.1 scanf()函数的格式 104
4.2.2 scanf()函数的格式字符 106
4.2.3 scanf()函数的注意事项 107
4.2.4 scanf()函数的返回值 109
4.3 其他常用输入/输出函数 110
4.3.1 getchar()函数 110
4.3.2 getch()函数 111
4.3.3 gets()函数 111
4.3.4 putch()函数 113
4.3.5 puts()函数 113
4.4 拓展训练 114
4.4.1 训练一：鸡、兔同笼问题 114
4.4.2 训练二：格式化输入 115
4.4.3 训练三：输入两个变量的值，交换输出 116
4.4.4 训练四：大小写字母转换器 117
4.5 技术解惑 118
4.5.1 ASCII 码查询 118
4.5.2 输入/输出的概念 118

第 5 章 程序的三大流程控制结构 120
5.1 顺序结构 120
5.2 条件语句 120
5.2.1 if 语句 121
5.2.2 if-else 语句 124
5.2.3 if-else-if 语句 125
5.2.4 if 语句的嵌套 126
5.2.5 布尔表达式 131
5.2.6 增强后的 switch 分支语句 131
5.3 循环语句 134
5.3.1 while 循环结构 134
5.3.2 do-while 循环结构 141
5.3.3 for 循环结构 144
5.3.4 循环的嵌套 150
5.4 转向语句 157
5.4.1 break 中断语句 157
5.4.2 continue 条件继续语句 159
5.4.3 标签语句 161
5.5 返回语句 162
5.6 拓展训练 163
5.6.1 训练一：水仙花数 163
5.6.2 训练二：工资提成问题 164
5.6.3 训练三：学生成绩评分等级 165
5.6.4 训练四：商店找零问题 166
5.6.5 训练五：斐波那契数列问题 167
5.6.6 训练六：简单计算器 168
5.7 技术解惑 169
5.7.1 判断一个数是否为素数 169
5.7.2 判断三角形构成问题 170

第 2 篇 核心技术篇

第 6 章 数组管理更复杂的数据 171
6.1 了解数组 171
6.1.1 使用数组的好处 171

6.1.2 数组的概念.....173
6.1.3 数组的维数.....173
6.2 一维数组.....174
6.2.1 一维数组的声明.....174
6.2.2 一维数组的存储.....177
6.2.3 引用一维数组.....178
6.2.4 一维数组的初始化.....180
6.3 二维数组.....182
6.3.1 二维数组的声明.....182
6.3.2 二维数组的存储.....183
6.3.3 二维数组的初始化.....186
6.4 字符数组和字符串.....188
6.4.1 字符数组.....188
6.4.2 了解字符串.....189
6.4.3 字符串的输入/输出.....191
6.5 拓展训练.....192
6.5.1 训练一：通过数组求解斐波那契数列问题.....192
6.5.2 训练二：采用冒泡排序法对数据进行排序.....193
6.5.3 训练三：通过数组计算学生的平均成绩.....194
6.5.4 训练四：字符串反转.....196
第 7 章 用函数模块化代码结构.....198
7.1 函数概述.....198
7.1.1 函数的概念.....198
7.1.2 函数的分类.....199
7.1.3 定义函数.....200
7.1.4 main()函数.....201
7.2 函数的工作过程.....202
7.2.1 程序结构.....202
7.2.2 函数执行过程.....203
7.3 编写函数.....204
7.3.1 函数头.....204
7.3.2 返回类型.....204
7.3.3 参数列表.....205
7.3.4 函数体.....206
7.3.5 函数原型.....206

7.4 函数的参数.....208
7.4.1 参数传递过程.....208
7.4.2 值调用.....209
7.4.3 引用调用.....210
7.4.4 数组调用.....213
7.4.5 main()函数的参数.....215
7.5 函数调用.....218
7.5.1 函数调用方式.....218
7.5.2 被调函数的说明.....219
7.5.3 返回函数结果.....220
7.5.4 函数的嵌套调用.....221
7.6 递归函数.....223
7.6.1 函数的递归调用.....223
7.6.2 递归的基本原理.....225
7.6.3 递归函数设计.....226
7.6.4 递归的优缺点.....229
7.7 拓展训练.....229
7.7.1 训练一：计算 1 到 x 的 5 次方累加和.....229
7.7.2 训练二：统计字符出现次数.....230
7.7.3 训练三：通过函数计算两个数的差值.....231
7.7.4 训练四：十进制数转换为二进制数.....232
7.8 技术解惑.....233
7.8.1 函数传址方式.....233
7.8.2 形参和实参.....236
第 8 章 算法在 C 语言中的应用.....237
8.1 算法的定义.....237
8.1.1 我所理解的算法.....237
8.1.2 什么情况下使用算法.....238
8.2 常见算法.....239
8.2.1 冒泡排序法.....239
8.2.2 折半查找.....240
8.2.3 堆栈的实现.....241
8.2.4 插入排序法.....244
8.3 拓展训练.....245
8.3.1 训练一：求最大值算法.....245
8.3.2 训练二：希尔排序.....246

8.4 技术解惑.....248
8.4.1 算法的必要性.....248
8.4.2 当下流行的各种算法.....248
第 9 章 用指针控制计算机的虚拟地址.....253
9.1 内存和变量.....253
9.1.1 计算机内存.....253
9.1.2 变量的存储.....254
9.2 指针和简单变量.....256
9.2.1 指针的概念.....256
9.2.2 创建指针.....257
9.2.3 初始化指针变量.....258
9.2.4 指针变量的引用.....259
9.2.5 给函数传递指针.....263
9.3 指针变量的赋值.....264
9.3.1 初始化赋值.....264
9.3.2 取地址赋值.....265
9.3.3 指针之间赋值.....265
9.3.4 数组赋值.....266
9.3.5 字符串赋值.....266
9.3.6 函数入口赋值.....269
9.4 指针和数组的关系.....270
9.4.1 指针、数组和地址之间的关系.....270
9.4.2 指针变量的运算.....271
9.4.3 用指针操作数组元素.....273
9.5 指向多维数组的指针.....276
9.5.1 理解二维数组的地址.....276
9.5.2 多维数组的指针表示.....279
9.5.3 指向多维数组的指针变量.....280
9.5.4 数组名作为函数的参数.....284
9.5.5 指向数组的指针小结.....288
9.6 指针和字符串.....288
9.6.1 字符串的指针表示.....289
9.6.2 字符串指针作为函数参数.....290
9.6.3 字符数组和字符串指针变量的区别.....292
9.7 指针数组.....294
9.7.1 指针数组的概念.....294

9.7.2 用指针数组处理字符串 296
9.7.3 用指针数组作为函数参数 298
9.8 指向指针的指针 300
9.8.1 理解指向指针的指针 300
9.8.2 二级指针变量与数组 303
9.9 指针和函数 307
9.9.1 返回指针的函数 307
9.9.2 指向函数的指针 309
9.10 指针和 const 变量 312
9.10.1 用 const 控制指针 312
9.10.2 const 的几种特殊用法 313
9.11 拓展训练 314
9.11.1 训练一：学生成绩管理 314
9.11.2 训练二：输出两个数中的大者 315
9.11.3 训练三：通过指针变量交换两个数 316
9.12 技术解惑 318
9.12.1 指向指针的指针有什么用处 318
9.12.2 数组指针和指针数组的区别 319
第 10 章 不同形式的存储——结构、联合和枚举 321
10.1 结构的概念 321
10.1.1 定义结构类型 321
10.1.2 定义结构变量 323
10.1.3 使用结构变量 323
10.2 结构的嵌套 325
10.2.1 包含数组的结构 326
10.2.2 包含指针的结构 326
10.2.3 包含结构的结构 327
10.3 初始化结构变量 330
10.4 结构数组 332
10.4.1 结构数组的定义和引用 332
10.4.2 结构数组的初始化 333
10.4.3 结构数组实例 334
10.5 结构指针 335
10.5.1 定义结构指针 335
10.5.2 使用结构指针 336
10.5.3 用指针处理结构数组 336

10.6 向函数传递结构.....338
10.6.1 传递结构变量的值到函数.....338
10.6.2 传递结构指针到函数.....339
10.7 联合.....340
10.7.1 结构和联合的区别.....340
10.7.2 定义联合类型.....341
10.7.3 使用联合变量.....342
10.7.4 在结构中嵌套联合类型.....342
10.7.5 联合数组.....344
10.7.6 联合指针.....345
10.8 枚举.....346
10.8.1 定义枚举类型.....346
10.8.2 使用枚举变量.....347
10.8.3 枚举类型变量的赋值.....347
10.9 使用 typedef.....348
10.10 拓展训练.....349
10.10.1 训练一：统计并输出学生和老师的信息.....349
10.10.2 训练二：结构变量地址作为函数参数实例.....350
10.10.3 训练三：候选人票数统计.....352
10.11 技术解惑.....353
10.11.1 结构变量地址作为函数参数.....353
10.11.2 用户自定义类型.....353



21世纪经济管理新形态教材 · 统计学系列

商务数据挖掘与可视化
基于Python的应用

谢贤芬
古万荣 ◎ 主 编

新形态教材

课程思政特色教材

清华大学出版社

内 容 简 介

本书以商务数据处理与应用为基本出发点，重点介绍基于 Python 语言的商务数据处理技巧、数据分析方法及可视化操作，内容包括：常见的数据管理及分析软件介绍，Python 编程基础，Matplotlib、Seaborn 和 Pyecharts 数据可视化，线性相关与回归分析，数据压缩、聚类、时间序列、大数据分析方法，综合案例实战。

本书案例内容丰富，通过 26 个综合案例详细介绍了数据挖掘方法及可视化在商务数据中的应用，展现了 Python 的强大数据分析能力。读者在了解数据挖掘方法原理的基础上，通过案例学习可以更有效地提高其数据分析的实际能力。本书适用于各个层次的读者，既可面向非计算机专业的数据分析者，强调应用性和实战性，也可作为计算机专业或数据科学专业的基础应用学习书。

本书封面贴有清华大学出版社防伪标签，无标签者不得销售。
版权所有，侵权必究。举报：010-62782989，beiqinquan@tup.tsinghua.edu.cn

图书在版编目（CIP）数据

商务数据挖掘与可视化：基于 Python 的应用 / 谢贤芬，古万荣主编. —北京：清华大学出版社，2023.12

21 世纪经济管理新形态教材. 统计学系列

ISBN 978-7-302-63660-1

Ⅰ. ①商… Ⅱ. ①谢… ②古… Ⅲ. ①商业信息－数据采集－教材 Ⅳ. ①F713.51

中国国家版本馆 CIP 数据核字(2023)第 100101 号

责任编辑：胡　月
封面设计：汉风唐韵
责任校对：王荣静
责任印制：沈　露
出版发行：清华大学出版社

网　　址：https://www.tup.com.cn，https://www.wqxuetang.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-83470000　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课 件 下 载：https://www.tup.com.cn，010-83470332

印 装 者：三河市龙大印装有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　张：**22.75　　**字　　数：**518 千字
版　　次：2023 年 12 月第 1 版　　**印　　次：**2023 年 12 月第 1 次印刷
定　　价：69.00 元

产品编号：096542-01

目录

第 1 章　数据管理及分析软件 9
1.1　数据收集及管理 9
1.1.1　数据的类型 9
1.1.2　数据的收集 10
1.1.3　数据的管理 12
1.2　数据挖掘软件 13
1.2.1　常见统计分析软件 13
1.2.2　Python 语言 14
1.3　习题 15
第 2 章　Python 编程基础 16
2.1　Python 简介 16
2.2　数据挖掘编程平台 18
2.2.1　Anaconda 科学计算发行包 18
2.2.2　PyCharm 分析平台 21
2.2.3　Jupyter 编辑平台 22
2.3　Python 基本语法 25
2.3.1　基本数据类型 25
2.3.2　基本语句 31
2.3.3　函数的定义与使用 35
2.3.4　案例：网络电影的播放量排行榜 37
2.4　基本分析包 40
2.4.1　数值分析包 numpy 40
2.4.2　数据分析包 pandas 42
2.4.3　基本描述统计分析函数 46
2.5　习题 49
第 3 章　Matplotlib 数据可视化 51
3.1　Matplotlib 基本参数配置 51
3.1.1　线条的设置 51
3.1.2　坐标轴的设置 53
3.1.3　图例的设置 55
3.2　绘图参数文件及主要函数 57
3.2.1　修改绘图参数文件 57
3.2.2　绘图主要函数 61
3.3　Matplotlib 基本绘图 64
3.3.1　直方图的绘制及参数说明 64
3.3.2　折线图的绘制及参数说明 66
3.3.3　饼图的绘制及参数说明 68
3.3.4　散点图的绘制及参数说明 72
3.3.5　箱线图的绘制及参数说明 75
3.3.6　其他图表的绘制 78
3.4　Matplotlib 绘图案例 84
3.4.1　案例：直方图分析企业员工业绩 84
3.4.2　案例：折线图分析商品销售情况 86

3.4.3 案例：基于饼图比较商品销售额 87
3.4.4 案例：散点图分析员工收入与学历的关系 89
3.4.5 案例：利用箱线图分析销售经理业绩 90
3.4.6 案例：利用面积图分析人口情况 91
3.5 习题 93
第 4 章 Seaborn 数据可视化 94
4.1 Seaborn 库简介 94
4.2 基本绘图参数设置 94
4.2.1 分组绘图设置 94
4.2.2 分面绘图设置 99
4.3 统计绘图函数 102
4.3.1 因子变量绘图 102
4.3.2 数值变量绘图 108
4.3.3 两变量关系绘图（jointplot） 112
4.3.4 时间序列图（tsplot） 114
4.3.5 热力图（heatmap） 115
4.4 Seaborn 可视化案例 117
4.4.1 案例：粽子销售数据的可视化 117
4.4.2 案例：基于 Python 的招聘信息可视化 119
4.4.3 案例：GDP 增长率的走势分析 122
4.5 习题 125
第 5 章 Pyecharts 数据可视化 126
5.1 Pyecharts 全局参数配置 126
5.1.1 基本元素配置 127
5.1.2 坐标轴配置 136
5.2 数据系列配置 139
5.2.1 样式类配置 140
5.2.2 标记类配置 146
5.2.3 其他配置 151
5.3 运行环境 154
5.3.1 生成 HTML 154
5.3.2 生成图片 154
5.3.3 案例：基于地图分析法的经济发展差异分析 156
5.4 习题 160
第 6 章 线性相关与回归分析 162
6.1 单变量线性相关与回归 162
6.1.1 单变量线性相关模型 162
6.1.2 单变量线性回归模型 163
6.1.3 案例：我国 CBA 球员得分能力的影响因素分析 165
6.2 多元线性相关与回归 168
6.2.1 多元线性相关模型 168
6.2.2 多元线性回归模型 169
6.2.3 案例：我国 CBA 球员得分能力的影响因素分析续 172
6.3 综合案例：我国财政收入水平的区域差异分析 174
6.3.1 数据来源及说明 175

6.3.2 描述性统计分析 175
6.3.3 相关性分析 176
6.4 习题 179
第 7 章 数据压缩分析方法 181
7.1 主成分分析的概念 181
7.1.1 基本思想 181
7.1.2 主成分的推导 182
7.2 主成分的基本分析 186
7.2.1 主成分分析的基本步骤 186
7.2.2 主成分的计算 187
7.2.3 主成分图 188
7.3 综合案例 192
7.3.1 案例：综艺节目分类分析 192
7.3.2 案例：中国城镇居民生活消费分布规律分析 194
7.4 习题 198
第 8 章 聚类分析方法 200
8.1 聚类分析的概念 200
8.1.1 基本思想 200
8.1.2 聚类统计量 200
8.2 常见的聚类分析方法 203
8.2.1 系统聚类法 203
8.2.2 快速聚类法 204
8.2.3 有序聚类法 205
8.2.4 模糊聚类法 206
8.3 综合案例 209
8.3.1 案例：经济指数的政府微博发展水平分析 210
8.3.2 案例：综艺节目分类的分析 215
8.3.3 案例：中国城镇居民生活消费分布规律分析 218
8.4 习题 220
第 9 章 时间序列分析方法 222
9.1 时间序列的概念 222
9.1.1 基本含义 222
9.1.2 应用领域 223
9.2 时间序列基本模型 223
9.2.1 AR 模型 223
9.2.2 MA 模型 225
9.2.3 ARMA 模型 226
9.2.4 ARIMA 模型 226
9.3 ARIMA 模型的构建 227
9.3.1 基本步骤 227
9.3.2 序列的平稳性检验 228
9.3.3 模型的识别与定阶 230
9.3.4 模型的检验 234
9.3.5 数据预测 236
9.3.6 案例：上证综合指数走势的预测分析 236

9.3.7 案例：流感数据的预测分析 240
9.4 综合案例：社会消费品零售总额分析 243
9.5 习题 253
第 10 章 大数据分析方法 254
10.1 大数据的概念及应用 254
10.1.1 基本含义 254
10.1.2 应用领域 254
10.2 文本挖掘基础 255
10.2.1 Python 文本预处理 255
10.2.2 绘制词云图 258
10.2.3 案例:基于 B 站视频弹幕的文本挖掘分析 260
10.3 网络爬虫及应用 262
10.3.1 爬虫的基本概念 262
10.3.2 常见网络爬虫框架 262
10.3.3 网络爬虫的应用领域 266
10.3.4 案例:基于 Python 爬虫的豆瓣华语电影数据可视化分析 266
10.4 习题 276
第 11 章 综合案例实战 277
11.1 概述 277
11.1.1 数据挖掘与可视化总结 277
11.1.2 商业智能与数据挖掘 277
11.1.3 SMARTBI 简介及其使用 278
11.2 基于国内旅游情况数据的可视化分析 281
11.2.1 案例介绍 281
11.2.2 基于 Python 实现 282
11.2.3 基于 SMARTBI 实现 288
11.3 基于某地二手房房价的可视化分析 292
11.2.1 案例介绍 292
11.2.2 基于 python 实现 293
11.2.3 基于 SMARTBI 实现 299
11.4 基于热销私家车销量数据的可视化分析 304
11.4.1 案例介绍 304
11.4.2 基于 python 实现 305
11.4.3 基于 SMARTBI 实现 310

13719398640.

教育部司局函件

教社科司函〔2018〕154号

2018年度教育部人文社会科学研究一般项目立项通知书

华南农业大学古万荣同志：

您申报的《**基于超图模型的新闻推荐及应用研究**》项目，经我部组织专家评审并经公示，正式批准为2018年度教育部人文社会科学研究青年基金项目。

项目批准号：**18YJCZH037**

批准经费：**8** 万元。请按照《高等学校哲学社会科学繁荣计划专项资金管理办法》（财教〔2016〕317号）以及国家有关科研经费管理制度，合理合规使用经费，提高经费使用效益。

立项时间：**2018年7月24日**。项目研究周期一般为3年。我部将于2020年统一组织项目中期检查，年度预算执行情况是项目中期检查的内容之一。中期检查情况将作为后续拨款的重要依据。

请按照《教育部人文社会科学研究项目管理办法》（教社科〔2006〕2号）的要求和您提交的申请评审书中设计的研究内容及研究计划开展项目研究，确保项目按期保质保量完成。项目鉴定、结项按照《教育部人文社会科学研究项目成果鉴定和结项办法》（教社科司函〔2007〕145号）进行。所有出版或发表的项目研究成果，须在显著位置标明“教育部人文社会科学研究规划基金/青年基金/自筹经费项目”字样和项目批准号，否则项目中期检查及鉴定结项不予通过。

教育部社会科学司

2018年7月24日

广东省哲学社会科学规划领导小组办公室

粤社科规划办通[2018]31号

广东省哲学社会科学“十三五”规划2018年度资助项目立项通知书

古万荣 同志：

经专家会议评审及省哲学社会科学规划领导小组审批，您申报的广东省哲学社会科学“十三五”规划2018年度一般项目《基于复杂对象关系模型的新闻推荐研究及应用》获准立项，批准号：GD18CXW01，资助经费5万元，第一次拨款4.5万，预留经费0.5万。请认真填写立项回执，并按时交回我办。

填写立项回执后，项目申请书即成为立项协议，对项目负责人及其所在单位具有约束力。项目负责人所在单位须对立项协议的履行承担保证责任。项目负责人及所在单位须了解和执行以下规定：

1. 课题组须学习并遵守《广东省哲学社会科学规划项目管理办法》。对于出现违规行为的，我办依据《广东省哲学社会科学规划项目管理办法》的相关规定进行处理。

2. 立项项目经费不再追加。课题组如不接受，我办将撤消该项目立项。而立项经费一经接受，课题组将不得以资助经费不足为由，

擅自变更原设计的最终成果形式和内容。

3、项目经费分两次拨付，其中，预留经费将在项目通过鉴定结项后拨付。拨款帐号、开户行、户名如有变动，须以书面形式及时告知我办，以便准确拨款。

4、项目研究过程中，如有变更项目负责人、延长完成时间、改变成果内容或形式、变更项目管理单位、变更或增补课题组成员、中止项目或撤消项目等重要事项，项目负责人或所在单位必须按要求填写《广东省哲学社会科学规划项目重要事项变更审批表》，及时我办审批。

5、项目成果的鉴定结项由我办组织，实行匿名鉴定制度。项目鉴定结项所需材料及装印要求，详见“广东社科规划”网站。

6、成果鉴定等级分为优秀、良好、合格和不合格四个等次。不合格者即未能通过结项，不予拨付预留经费。项目成果的最终鉴定等级均通过“广东社科规划”网站予以公示，并通知项目负责人所在单位。

7、为提高成果质量，计划出版的成果必须在鉴定结项后方可出版。对于违反规定擅自出版的，视为项目负责人自行终止相关资助协议，我办将不受理结项申请，并按有关规定扣留并追回项目研究经费。

以上规定，项目负责人及所在单位应严格遵守。如有异议，可不接受资助（在回执中注明），立项协议自行废止。

联系人：钟月仙、尹 梅

电　话：（020）83825078

地　址：广州市天河北路618号广东社科中心B座920室

邮　编：510635

广东省哲学社会科学规划领导小组办公室

2018年11月9日

29

受理编号：c18140500000381

项目编号：2018A030313437

文件编号：粤科规财字〔2018〕0105号

（广东科技微信公众号）

（受理纸质材料二维码）

广东省自然科学基金项目

合同书

项目名称：协同超图推荐模型研究及应用

项目类别：广东省自然科学基金-自由申请

项目起止时间：2018-05-01 至 2021-04-30

管理单位（甲方）：广东省自然科学基金管理委员会

依托单位（乙方）：华南农业大学

通讯地址：广东省广州市天河区五山路483号

邮政编码：510642　　单位电话：020-38632819

项目负责人：古万荣　　联系电话：13719398640

广东省科学技术厅

二〇一四年制

一、主要研究内容和要达到的目标

主要研究内容：

（1）研究推荐算法中，各异构对象及其相互关系如何映射为超图模型节点、超边和权重。该研究内容将研究如何将传统的推荐算法中的二分图关系映射为多元异构对象关系，并用超图模型描述其多元关系，进一步地，需要研究多元异构对象关系之间的权重和关联类别等。

（2）研究超图模型中多元异构节点的快速聚类方法。传统方法注重于同构对象聚类，比如基于内容的物品聚类等，本研究拟通过超图集合模型，通过带权重的超边关联进行多元异构对象聚类。从而实现挖掘推荐技术中用户和物品之间的兴趣关系。

（3）研究超图模型的节点排序算法。在推荐技术场景中，给定一个用户节点，如果能给出待推荐物品节点对这个用户节点的超空间距离，则可以给出推荐结果列表。本研究拟基于超图模型，给出给定节点的排序算法，以快速有效地锁定待推荐物品。

（4）将超图推荐模型应用到新闻推荐中，并研究新闻推荐中新闻增量推荐问题。

研究目标：

（1）建立推荐系统的超图模型框架。本研究将传统模型中的基于内容推荐和基于协同过滤推荐的方法融合到超图模型中，通过对节点和超边的定义来调整融合模型。

（2）本研究将解决在超图模型中海量多元异构节点快速聚类问题，研发节点增量自适应超图聚类方法。

（3）本研究将解决在超图模型中的排序问题，重点解决超图模型在推荐应用中的快速排序问题、节点和超边增量排序调整问题。

（4）本研究将协同超图推荐模型应用到新闻推荐中，通过超图模型参数接口自适应调整新闻推荐模型的内容和协同性参与因素，形成研究模型的应用示范。

二、研究成果及形式

<table>
<tr><td rowspan="2">论文及专著情况</td><td colspan="2">国家统计源刊物以上刊物发表论文（篇）</td><td colspan="2">3</td><td colspan="2" rowspan="2">科技报告（篇）</td><td colspan="2" rowspan="2">1</td></tr>
<tr><td colspan="2">专著（册）</td><td colspan="2">0</td></tr>
<tr><td rowspan="3">专利情况(项)</td><td colspan="2">发明专利</td><td colspan="2">实用新型专利</td><td colspan="2">外观设计专利</td><td colspan="2">国外专利</td></tr>
<tr><td>申请</td><td>授权</td><td>申请</td><td>授权</td><td>申请</td><td>授权</td><td>申请</td><td>授权</td></tr>
<tr><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>其他</td><td colspan="8"></td></tr>
</table>

三、项目进度和阶段目标

1. 项目起止时间： 2018-05-01 至 2021-04-30

2. 项目实施进度及阶段主要目标：

开始日期	结束日期	主要工作内容
2018-05-01	2019-04-30	做好实验数据收集与预处理工作，着重开展本项目的理论研究，并为应用研究进行准备工作。在文献充分检索的基础上，结合超图理论、主成分分析、异构对象聚类、谱聚类、聚类关联分析，研究面向多异构对象超图聚类的推荐模型，着重研究解决推荐模型主要内容对象挖掘、超图模型定义和构建、超图聚类、用户感兴趣的物品选择优化等问题。预期发表学术论文1篇。参加国际学术会议1人次。参加国内学术会议1人次。
2019-05-01	2020-04-30	在文献充分检索的基础上，继续完善基于超图聚类的推荐模型研究，并研究具备基于超图排序的协同推荐模型。基于超图模型表达，研究超图模型中的排序最优化问题，进而建立多异构对象排序的推荐新算法。进一步地，研究所建立的基于超图聚类和排序的推荐及应用，并研究超图模型构建的自适应方法。预期发表学术论文1篇，申请发明专利1项。参加国内学术会议2人次。
2020-05-01	2021-04-30	在前两年研究工作的基础上，进一步深入基于超图模型的推荐模型研究，着重研究基于超图模型的应用研究，将通过在线新闻系统和数据进行新闻推荐研究，并研究超图模型在实际系统中的自适应问题。预期发表学术论文1篇。参加国内学术会议1人次。

四、项目总经费及省科技厅经费预算

1. 省科技厅经费下达总额： （大写）壹拾万圆整；（小写 ）10万元；					
2. 省科技厅经费年度下达计划：					
年度	2018 年	年	年	年	年
经费(万元)	10.00				
3. 总经费开支预算计划：					
经费筹集情况：					（单位：万元）
省科技厅经费	自筹资金				合计
	自有资金	贷款	地方政府投入	其它	
10.00				0	10.00
政府部门、境外资金及其他资金投入情况说明：					

经费预算				（单位：万元）
	总投入经费		省科技厅经费	
支出经费	经费额	用途说明	经费额	用途说明
基建费：				
1、直接费用：	8.50		8.50	
(1)设备费：				
(2)材料费：				
(3)测试化验加工外协费：				
(4)燃料动力费：				
（5）差旅费/会议费/国际合作与交流费：	3.00	参加国际会议的会务费、差旅住宿费用；参加国际会议一次，约1万元；国内异地学术会议4次，约0.5万元/次	3.00	参加国际会议的会务费、差旅住宿费用；参加国际会议一次，约1万元；国内异地学术会议4次，约0.5万元/次
（6）出版/文献/信息传播/知识产权事务费：	2.50	发表论文的版面费，申请发明专利的费用；购买文献资料，打印、复印费用	2.50	发表论文的版面费，申请发明专利的费用；购买文献资料，打印、复印费用
（7）劳务费：	3.00	参与人员的劳务费用，0.05万元/人月，约需2人，共30个月。	3.00	参与人员的劳务费用，0.05万元/人月，约需2人，共30个月。
（8）人员费：				
（9）专家咨询费：				
（10）直接费用其他支出：				
（11）科技金融服务体系其他费用：	0.00		0.00	
①信用评级补贴：				
②大赛场租：				
③特派员奖励与补贴：				
2、间接费用：	1.50		1.50	
（1）间接成本：				
（2）管理成本：	0.50	5%的管理费	0.50	5%的管理费
（3） 绩效支出：	1.00	团队成员绩效奖励	1.00	团队成员绩效奖励
合计：	10.00		10.00	

五、人员信息

项目负责人

姓名	证件号码	年龄	性别	职称	学历	在项目中承担的任务	所在单位	签名
古万荣	440229198201270018	36	男	讲师	博士研究生	项目负责人	华南农业大学	

项目组主要成员

姓名	证件号码	年龄	性别	职称	学历	在项目中承担的任务	所在单位	签名
田绪红	110108196609051473	52	男	教授	博士研究生	异构数据挖掘算法研究	华南农业大学	
李吉平	370627197103131133	47	男	教授	博士研究生	推荐系统的应用研究	华南农业大学	
王国华	441781198804130114	30	男	讲师	博士研究生	异构对象聚类和排序算法研究	华南农业大学	
韦婷婷	452623198610023349	32	女	讲师	博士研究生	超图模型随机游走算法设计与分析	华南农业大学	
黎嘉朗	440681199404134739	24	男	未取得	本科	系统搭建和应用实现	华南农业大学	
徐志文	34080319940925253X	24	男	未取得	本科	系统搭建和应用实现	华南农业大学	

六、依托单位与合作单位的合作协议

承担/参与单位名称（盖章）	工作分工	总经费分摊（万元）	省科技厅经费分配（万元）
华南农业大学	总体设计/进度把控和跟进 2万； 资料查询及调研 2万； 异构对象模型及其挖掘方法研究 2万； 算法设计与实验程序编写 2万； 资料分析和结果比较 2万。	10.00	10.00
南方报业传媒集团			
	合计	10.00	10.00

七、合同条款

第一条 甲方与乙方根据《中华人民共和国合同法》及国家有关法规和规定，为顺利完成（2018）年**协同超图推荐模型研究及应用** 专项项目（文件编号： **粤科规财字〔2018〕0105号**）经协商一致，特订立本合同，作为甲乙双方在项目实施管理过程中共同遵守的依据。

第二条 甲方的权利义务：

1. 按合同书规定进行经费核拨的有关工作协调。
2. 根据甲方需要，在不影响乙方工作的前提下，定期或不定期对乙方项目的实施情况和经费使用情况进行检查或抽查。
3. 根据《广东省科技计划项目信用管理办法(试行)》对乙方进行科技计划信用管理。

第三条 乙方的权利义务：

1. 确保落实自筹经费及有关保障条件。
2. 按合同书规定，对甲方核拨的经费实行专款专用，单独列账，并随时配合甲方进行监督检查。
3. 使用财政资金采购设备、原材料等，按照《广东省实施〈中华人民共和国招标投标法〉办法》有关规定，符合招标条件的须进行招标。
4. 项目实施完成或实施到一定程度，须按照《广东省省级科技计划项目结题管理的实施细则（试行）》提出验收或终止结题的申请，并按甲方要求做好项目结题工作。
5. 在每年规定时间内向甲方如实提交上年度工作情况报告，报告内容包含上年度项目进展情况、经费决算和取得的成果等。
6. 按照国家和省有关规定，提交科技报告及其他材料。

第四条 在履行本合同的过程中，如出现广东省相关政策法规重大改变等不可抗力情况，甲方有权对所核拨经费的数量和时间进行相应调整。

第五条 对分年度拨款（滚动资助）项目，甲方有权利根据项目研究进展或中期考核情况变更或中止项目后续资助经费数额。

第六条 在履行本合同的过程中，当事人一方发现可能导致项目整体或部分失败的情形时，应及时通知另一方，并采取适当措施减少损失，没有及时通知并采取适当措施，致使损失扩大的，应当就扩大的损失承担责任。

第七条 本项目技术成果的归属、转让和实施技术成果所产生的经济利益的分享，除双方另有约定外，按国家和广东省有关法规执行。

第八条 根据项目具体情况，经双方另行协商订立的附加条款，作为本合同正式内容的一部分，与本合同具有同等效力。

第九条 本合同一式三份，各份具有同等效力。甲、乙方及课题负责人各执一份，三方签字、盖章后即生效，有效期至项目结题后一年内。各方均应负合同的法律责任，不应受机构、人事变动的影响。

第十条 乙方必须接受甲方聘请的本项目合同监理单位的监督和管理。监理单位按照甲方赋予的权利对本项目合同的履行进行审核、进度调查，对项目合同变更、经费使用情况进行监督管理及组织项目验收。

说明：1. 本合同书中，凡是当事人约定无需填写的内容，应在空白处划（/）。

2. 委托代理人签订本合同书的，应出具合法、有效的委托书。

八、本合同签约各方

管理单位（甲方）：广东省自然科学基金管理委员会 （盖章）

法定代表人（或法人代理）：_____（签章）

年 月 日

依托单位（乙方）：华南农业大学 （盖章）

法定代表人（或法人代理）：陈晓阳 _____（签章）

联系人（项目主管）姓名：郑鹏 _____（签章）

Email：kjcgxk@scau.edu.cn

电话：020-85283435 / 13560344902

开户单位名称：华南农业大学

开户银行名称：广东广州工行五山支行

开户银行帐号：3602002609000310520

2018年 9月 6日

年 月 日

联系人（课题负责人）姓名：古万荣 （签名）

Email：guwanrong@scau.edu.cn

电话：13719398640

2018 年 8月27日

国 家 统 计 局

2020年度全国统计科学研究项目
立项通知书

古万荣 同志：

经国家统计局全国统计科学研究组织管理办公室（以下简称“科研管理办公室”）组织评审，您申请的全国统计科学研究项目（复杂多维数据的时空监控方法研究及其在经济社会中的应用）获准立项，立项编号：2020LY018，项目类别：一般项目。其中，重大项目资助经费总额15万元/项；重点项目资助经费总额5万元/项；一般项目自筹经费。资助项目的经费分期拨付，首期经费为总资助经费的80%，其余经费待结项后拨付。依据《全国统计科学研究项目管理修订办法》的规定，重大、重点类项目承担单位可按不低于1:1的比例配套科研经费。

为确保科研任务的顺利完成，现将有关事项通知如下：

一、项目实施时限为2020年9月15日至2022年9月14日。您及项目承担单位要认真组织实施，按时完成项目研究，并及时提交结项材料。提前完成的，可以申请提前结项。**研究项目需要延长期限的，由项目负责人在期限届满三个月前向科研管理办公室提交书面申请，延长期限不得超过半年。**

二、资助类项目请按照《全国统计科学研究项目资金管理修订办法》（详见中国统计科学研究网）的有关规定，根据批准的资助经费总额，科学合理、实事求是地编制项目经费预算，并在收到立项通知书之日起15日内填写项目经费预算审批表。无特殊情况，逾期

不提交的，视为自动放弃资助。预算审批表可从中国统计科学研究网的“科研管理”栏目下载。请严格执行国家有关科研经费管理制度，依法依规使用项目经费。

三、项目管理实行全过程跟踪管理制度。资助类项目负责人应在收到立项通知书后及时组织开展开题汇报，按照要求填写全国统计科学研究项目开题报告，并于30日内报送至科研管理办公室。开题报告请从中国统计科学研究网的“科研管理”栏目下载。

四、项目执行过程中实行重要事项报告制度。如有变更，由项目负责人填写研究项目变更申请表，经项目负责人所在单位审核并加盖公章后报科研管理办公室批准。

五、科研管理办公室负责组织实施资助类项目的结项鉴定。一般项目自行组织鉴定，科研管理办公室对是否符合结项要求进行审查。完成结项的项目颁发“全国统计科学研究项目结项证书”。**结项时，请登录中国统计科学研究网的“全国统计科学研究项目网络管理平台”查看结项注意事项并上传结项材料。**

六、有下列情形之一的，由科研管理办公室终止或撤销研究项目：1.研究成果（包括最终研究成果和阶段性研究成果）有严重政治问题的；2.研究成果存在知识产权、署名权等方面争议且争议未解决的；3.研究成果学术质量低劣的；4.剽窃他人成果的；5.与批准立项的研究项目论证材料严重不符的；6.逾期不提交延期申请，或延期到期仍不能完成的；7.严重违反财务制度的；8.其他。**被终止或撤销研究项目的资助经费应当追回。被撤销研究项目的项目负责人5年内不得申请研究项目。**

以上规定，项目负责人及责任单位应严格遵守。如有异议，可以不接受资助，立项协议自行废止。

国家统计局

2020年9月15日

国 家 统 计 局

2023年度全国统计科学研究项目立项通知书

古万荣 同志:

经国家统计局全国统计科学研究组织管理办公室（以下简称“科研管理办公室”）组织评审，您申请的全国统计科学研究项目获准立项。

项目名称：基于深度学习及混频数据模型的流行病组合预测方法研究

立项编号：2023LY016

项目类别：优选项目

请登陆“全国统计科学研究管理平台”查询立项信息。其中，重大项目资助经费总额15万元/项；重点项目资助经费总额5万元/项；优选项目自筹经费。资助项目的经费分期拨付，首期经费为总资助经费的80%，其余经费待结项后拨付。依据《全国统计科学研究项目管理办法》的规定，重大、重点类项目承担单位可按不低于1:1的比例配套科研经费。

为确保科研任务的顺利完成，现将有关事项通知如下：

一、项目实施时限为2023年11月7日至2025年11月7日。您及项目责任单位要认真组织实施，按时完成项目研究，并及时提交结项材料。提前完成的，可以申请提前结项。研究项目需要延长期限的，由项目负责人在期限届满三个月前向科研管理办公室提交书面申请，延长期限不得超过半年。

二、资助类项目请按照《全国统计科学研究项目资金管理办法》（详见中国统计科学研究网）的有关规定，根据批准的资助经费总额，科学合理、实事求是地编制项目经费预算，并在收到立项通知

书之日起15日内，通过“全国统计科学研究管理平台”填报《项目经费预算审批表》。无特殊情况，逾期不提交的，视为自动放弃资助。

请严格执行国家有关科研经费管理制度，依法依规使用项目经费。

三、资助类项目负责人应在收到立项通知书后及时开展开题汇报，并在30日内通过“全国统计科学研究管理平台”完成填报。

四、项目执行过程中涉及的重要事项变更审批，按照《全国统计科学研究项目管理办法》相关规定执行。

五、科研管理办公室负责组织实施资助类项目的结项鉴定。优选项目自行组织鉴定，科研管理办公室对是否符合结项要求进行审查。申请结项时，请通过“全国统计科学研究管理平台”上传结项材料。

完成结项的项目颁发《全国统计科学研究项目结项证书》。

六、有下列情形之一的，由科研管理办公室终止或撤销研究项目：1.研究成果（包括最终研究成果和阶段性研究成果）有严重政治问题的；2.项目研究中有剽窃他人科研成果或者弄虚作假等学术不端行为的；3.研究成果存在知识产权、署名权等方面争议且争议未解决的；4.研究成果学术质量低劣，未能达到申请书目标的；5.逾期不提交延期申请，或延期到期仍不能完成的；6.严重违反财务制度的；7.存在其他严重情况的。

被终止项目的负责人3年内不得申请或者参与申请研究项目，被撤销项目的负责人5年内不得申请或者参与申请研究项目。科研管理办公室视情节轻重分别作出退回结余资金、退回结余资金和绩效支出、退回已拨资金处理。

以上规定，项目负责人及项目责任单位应严格遵守。如有异议，可以不接受资助，立项协议自行废止。

国家统计局

2025年11月7日

受理编号：221115096431345

项目编号：2022020301

文件编号：云科函〔2022〕88号

云浮市科技计划项目合同书

221115096431345

项目名称：	大数据智慧种养平台关键技术研究及其应用示范		
计划类别：	农业信息化智能化研究与应用		
项目起止时间：	2022-07-01至2024-06-30		
管理单位（甲方）：	云浮市科学技术局		
承担单位（乙方）：	云浮市物联网研究院有限公司		
通讯地址：	云浮市新兴县新城镇沿江北路66号筠诚国际广场金融中心10楼		
邮政编码：	527400	单位电话：	0766-2929555
项目负责人：	郭杰	联系电话：	18926172121
项目联系人：	黄晓宁	联系电话：	0766-2929911
乙方主管部门（丙方）：	新兴县科技局		

云浮市科学技术局

二零二二年制

一、项目实施内容

研发内容:
本研究拟对新兴县农户的种养产业状况进行数字化信息管理，主要构建智慧种植管理平台和智慧养殖管理平台，同时与物联网传感大数据采集系统、广东省动物防疫监督管理系统、县域综合资源管理系统、广东省新兴县农产品质量安全溯源平台以及全国畜禽养殖场直联直报系统对接。
拟解决的问题及技术路线:
1、智慧种植管理平台
建设智慧种植管理平台的目的是通过信息化技术提高监管部门对农户的管理水平，提高农户的服务质量和使农户生产过程降本增效。农户在种植过程中，依靠传统的记录方法会造成较大的劳动力支出，监管部门审查的工作量也繁琐耗时，在一定程度上既不能保障数据的准确性还增加了生产成本。如何实现乡村整体的农户产业数字化管理，让农户与监管部门提高管理效益，本研究内容拟开发面向农业局相关管理人员的中心管理平台、面向镇级别的网点管理员的网点服务小程序和面向农户的生产管理小程序。
2、智慧养殖管理平台（养禽版）
本研究结合养禽业务的实际需求来建设新的养殖管理系统，拟建设一套面向养鸡业务管理系统，包括养殖管理后台系统、智养殖（养禽版）小程序和供监管部门使用的新兴县鸡只养殖监管平台。本研究力求系统能帮助农户有效监管养殖情况，助监管部门实时监察农户的养殖状况，以提升乡村农产数字化管理的服务质量。
3、多源信息融合预测研究
为更方便统计管理各乡镇的农业农产数据，本研究拟将各管理平台的多源数据信息进行融合处理，并对农产的价格和农产市场供求趋势进行预测，将预测信息推送到种养管理平台，使其既能按县—镇—村维度展示各种农产品的数据信息，又能统计分析新兴县的种养趋势。
创新点:
1、多源数据的融合问题。把不同来源、格式、性质的数据在逻辑上或物理上有机地集中，通过一种一致的、精确的、可用的表示法，对同一种现实世界中的实体对象的不同数据做整合。这将涉及到每个数据源的数据语义并解决歧义性等问题。
2、多源数据的联合分析与挖掘问题。本项目将研究多层次多源数据关联分析和挖掘框架，研究并设计新的多源数据的联合分析与挖掘模型，使用深度学习算法生成更多关联模式，从而解决数据关联和属性缺失等带来的数据稀疏性问题。
3、建立动态增量种养大数据挖掘模式。传统的农牧大数据挖掘与分析大多基于静态的批量数据挖掘，本项目拟在批量数据分析和挖掘的基础上，增加对流式增量数据的挖掘，研发自适应的种养大数据流式增量分析与挖掘范型。

2211150004313

二、项目考核指标

（一）项目完成后提供的研究开发成果及形式（须明确产品、专利、版权、标准等成果的类型及数量）

成果形式		成果数量	成果形式		成果数量
发明专利	申请	2	引进人才（人）		0
	授权		培养人才（人）		3
实用新型专利	申请	0	技术标准制定	牵头（个）	0
	授权			参与（个）	0
科技人才奖励（人）		0	科技报告（篇）		
外观专利	申请		软件著作权（项）		3
	授权		论文论著（篇）		0
获国家级奖项（项）		0	被收录论文数（篇）	SCI	0
获省级奖项（项）		0		IE	0
新产品（或新材料、新装备、新品种（系））		3		CA	0
国外专利	PCT受理		新工艺（或新方法、新模式、新技术）		0
	授权		新服务（项）		0
服务企业数量（家）		技术服务数量（项）			0
		服务企业数量（家）			0

2211150964931345

科技金融项目必填	开展培训宣讲活动场次（次）	
	服务企业数量（家）	
	帮助企业融资（万元）	
	引进专业机构（家）	
院士工作站项目必填	引进院士及其团队科技成果转化数量	0
	院士开展的战略咨询和技术指导次数	0
	院士年进站次数	0
	院士及院士团队年进站时间	
软科学项目必填	决策咨询报告（篇）（至少1篇）	
	研究总报告（篇）	
	研究中期报告（篇）	
	研究分报告（篇）	
	调研报告（篇）	
	专著（篇）[须注明“**市软科学研究计划项目（项目编号：）资助”]	
	培养人才（人）	
	获国家级奖项（项）	
	获省级奖项（项）	
	其他（具体形式：用户填）	

2211150964431345

其他成果及形式说明：

除以上实施绩效输出外，本项目还将研发种养系统原型及应用系统3套（智慧种植、智能养殖和智能预警管理平台），并将其应用到相关企业、养殖场或种植园，可以有效提高农业生产经营和管理的数字化、机械化和智能化水平，通过政企合作、数据共享、智能模块灵活拆解组装设计，推广平台和智能化模块的使用。

（二）主要技术经济指标及社会效益

累计新增销售收入（万元）	300
累计新增利税（万元）	75

其他主要技术经济指标及社会效益说明：

本项目完成后，将会产生大数据智慧种养平台原型及系统3套（智慧种植、智能养殖和智能预警管理平台），形成核心算法和专利2件以上。研究成果的产出和知识积累将给种植、养殖和管理部门带来全新的管理方式和预警预测方法，这些方法体系可以有效减轻种植和养殖联合管理过程中的冗余工作量，为融合数据分析、管理、预测和预警提供必不可少的核心组件和方法。项目成果将在多个种植园和养殖场应用。

项目负责人（签章）： 年 月 日

2211150943134345

三、项目进度和阶段目标

（一）项目起止时间：	2022-07-01 至 2024-06-30
（二）项目实施进度及阶段主要目标	
开始日期--结束日期	主要工作内容（限 500 字）
2022-07-01 至 2022-12-31	主要研究内容： 开展项目文献和数据整理，撰写基于多源数据融合的大数据种养系统分析与挖掘项目实施报告，同时对多源流式大数据进行深入研究，并实现多源大数据种养数据融合管理平台原型。 预期研究成果： （1）多源大数据种养数据融合管理平台原型实施方案； （2）申请1项软件著作权登记；
2023-01-01 至 2023-06-30	主要研究内容： 对多源数据融合进行深入研究，对各种传统系统和多源数据库场景进行研究，并实现系统数据级别和系统级别关联融合原型。 预期研究成果： （1）申请专利1 项； （2）申请1项软件著作权登记；
2023-07-01 至 2023-12-31	主要研究内容： 对多源数据联合分析和挖掘进行深入研究，对深度学习模型如GAN网络等进行深入研究，并实现多源数据联合分析和挖掘的技术原型。 预期研究成果： （1）申请1项软件著作权登记； （2）申请专利1项；
2024-01-01 至 2024-06-30	主要研究内容： 综合前面研究，理清相关方案中的各种细节问题，将各种分析原型和模型进行有机整合，实现高性能、低消耗和高准确度的大数据种养融合管理平台原型系统。 预期研究成果 （1）项目累计获得3项软件著作权登记； （2）项目累计申请专利2项； （3）项目累计培养研究生3名； （4）研发大数据种养融合系统一套，在3个种植/养殖场示范应用； （5）项目结题报告。

2211305643345

四、承担、参与单位工作分工及经费分配情况

承担/参与单位名称（盖章）	工作分工	总经费分摊（万元）	云浮市科学技术局经费分配（万元）
云浮市物联网研究院有限公司	项目研发牵头，提供数据场景和应用场景，联系种植场和养殖场进行应用示范，对系统开发和关键技术提供功能和性能需求规划。	125	25
华南农业大学	负责数据挖掘、机器学习和人工智能技术研究，完成大数据智慧种养系统设计和原型实现。	25	25

2211150964313345

五、项目总经费及云浮市科学技术局经费预算

（一）云浮市科学技术局经费下达总额：（大写） 伍拾万元整；（小写）：50 万元

（二）云浮市科学技术局经费拨付方式： 一次性拨款

（三）云浮市科学技术局经费经费年度下达计划： 第一期50万元

（四）总经费及云浮市科学技术局经费开支预算计划：

经费筹集情况：（单位：万元）

总投入经费：150万元

	云浮市科学技术局经费	自筹资金				合计
		自有资金	贷款	地方政府投入	其它	
已投入经费	0	0	0	0	0	0
新增经费	50	100	0	0	0	150

其它政府部门、境外资金及其他资金投入情况说明：

无

（二）新增经费预算（单位：万元）

	新增经费总额		**其中：云浮市科学技术局经费**	
支出经费	**经费额**	**用途说明**	**经费额**	**用途说明**
基建费	0	/		
1、直接费用	141	主要包括设备费、材料费、燃料动力费、差旅费、会议费、国际合作与交流费、出版版/文献/信息传播	41	主要包括设备费、材料费、燃料动力费、差旅费、会议费、国际合作与交流费、出版版/文献/信息传播/知识产权事务费、人员费、劳务费、专家费等

2211509643134 5

		/知识产权事务费、人员费、劳务费、专家费等		
（1）设备费	8	购置本项目研发所需的设备GPU高性能服务器、数据节点服务器等	8	购置本项目研发所需的设备GPU高性能服务器、数据节点服务器等
（2）材料费	64	项目相关研发，成果应用示范物料、监控摄像头、交换机、路由器等费用	8	项目相关研发，成果应用示范物料、监控摄像头、交换机、路由器等费用
（3）测试化验加工外协费	0	/	0	/
（4）燃料动力费	2	研发过程中研发设备所发生的水电动力能源费，示范应用基地实验所需水、电等，由申报单位自筹	0	研发过程中研发设备所发生的水电动力能源费，示范应用基地实验所需水、电等，由申报单位自筹
（5）差旅费/会议费/国际合作与交流费	3	相关差旅、会议及交流费用	3	相关差旅、会议及交流费用
（6）出版/文献/信息传播/知识产权事务费	5	项目相关出版费、资料费、专利费、通信费、专用软件购买费、文献检索费及其他知识产权费用	5	项目相关出版费、资料费、专利费、通信费、专用软件购买费、文献检索费及其他知识产权费用
（7）劳务费	15	项目参与	15	项目参与人员劳务费

2211159643245

		人员劳务费		
（8）人员费	42	参与项目人员工资费用	0	参与项目人员工资费用
（9）专家咨询费	2	项目研发过程中的技术难点咨询	2	项目研发过程中的技术难点咨询
（10）直接费其他支出	0	/	0	/
（11）科技金融服务体系其他费用	0	/	0	/
①信用评级补贴	0	/	0	/
②大赛场租	0	/	0	/
③特派员奖励与补贴	0	/	0	/
2、间接费用	9	项目管理成本、绩效支出	9	项目管理成本、绩效支出
（1）间接成本	0	/	0	/
（2）管理成本	7	项目管理费用支出	7	项目管理费用支出
（3）绩效支出	2	项目激励及绩效支出	2	项目激励及绩效支出
合计	150		50	

2211509643134 5

六、人员信息

项目负责人：								
姓名	性别	年龄	职务	职称	学历	在项目中承担的任务	所在单位	签名
郭杰	男	47	工程技术中心负责人	中级,经济师	硕士研究生	负责项目全面技术和管理工作，负责项目统筹实施，把控项目研发方向及进度，包括企业与高校。	云浮市物联网研究院有限公司	
古万荣	男	40	硕士生导师/软件工程教研室主任	讲师	博士研究生	项目全面技术和管理工作，负责项目统筹实施，把控项目研发方向及进度。	华南农业大学	
毛宜军	男	43	硕士生导师	讲师	博士研究生	深度学习模型研究与实现，多源数据集成、融合、分析与流式数据处理框架设计与实现。	华南农业大学	
主要研究开发人员：								
姓名	性别	年龄	职务	职称	学历	在项目中承担的任务	所在单位	签名
黄晓宁	男	40	研发管理员	助理工程师	本科	负责项目研发管理相关协调工作，完成项目示范应用工作。	云浮市物联网研究院有限公司	
罗国汉	男	39	软件工程师	高级工程师	本科	种养数据深度挖掘和分析	云浮市物联网研究院有限公司	
陈燕平	男	27	软件工程师	未取得	本科	物联网软件系统开发	云浮市物联网研究院有限公司	
伍晓仪	女	28	科研项目管理	未取得	本科	负责研发人员组织协调、协助项目进度管理等工作	云浮市物联网研究院有限公司	

刘心	女	45	教师	副教授	博士研究生	对深度学习模型入GAN等进行深入研究	华南农业大学	
王国华	男	34	教师	讲师	博士研究生	多源数据融合研究，对各种传统系统和多源数据库场景进行研究	华南农业大学	
崔金荣	女	38	教师	讲师	博士研究生	智慧养殖管理平台（养禽版）研究	华南农业大学	
黄栋	男	35	教师	副教授	博士研究生	多源数据融合研究，对各种传统系统和多源数据库场景进行研究	华南农业大学	
柯海萍	女	25	学生	未取得	本科	数据挖掘分析	华南农业大学	
陈梓明	男	24	学生	未取得	本科	数据挖掘分析	华南农业大学	
卢泽伦	男	25	学生	未取得	本科	数据挖掘分析	华南农业大学	
何亦琛	男	24	学生	未取得	本科	数据挖掘分析	华南农业大学	
何浩明	男	27	学生	未取得	本科	数据挖掘分析	华南农业大学	

2211150964313345

七、承担、参与单位合作协议（须与申报书中合作协议或意向书相一致）

联合申报项目协议书
项目申报单位（甲方）：云浮市物联网研究院有限公司
项目参与单位（乙方）：华南农业大学

甲乙各方就共同申报云浮市2022年省科技创新战略专项（“大专项＋任务清单”）中：农业信息化智能化研究与应用的“大数据智慧种养平台关键技术研究及其应用示范”项目的相关事宜，经友好协商，达成如下合作协议：

第一条 项目研究工作详细分工：
甲方（主持方）：提供技术开发物联网数据和应用场景数据，负责技术数据模型研究，提供软硬件实验环境和业务专家组等协调、组织和沟通。
乙方（参与方）：负责数据挖掘、机器学习和人工智能技术研究，完成大数据智慧种养平台系统设计和原型实现。

第二条 经费分配：
1.若本申报项目获批立项，甲方和乙方同意综合各级政府下达的资助经费，各自分别按5：5比例分配经费。
2.甲方在收到云浮市科学技术局资助经费后一个月内按照经费50%支付给乙方指定账户：
乙方账户：华南农业大学
乙方账号：3602002609000310520
开户行：中国工商银行广州市五山支行

第三条 甲、乙双方课题组负责人：甲方单位项目负责人郭杰，联系电话为18926172121;乙方单位项目负责人为古万荣，联系电话为13719398640。

第四条成果归属：
1.项目实施过程中所产生的知识产权，优先执行任务下达单位的知识产权管理政策，在此前提下，各方独立完成的所有权归各自所有；双方共同完成的由双方共享。
2.项目成果申报各级奖项，双方单位排名根据具体情况另行商定，人员排名原则上按贡献大小先后排名。
3.专利申请：在经过对方同意情况下，甲乙双方可以本方独立完成的部分申请专利；联合申请专利时，单位顺序根据情况商定，发明人顺序按照实际贡献大小先后排名。
4.若项目成果转化，收益分配具体按照双方的贡献大小进行分配或双方另行商定。

第五条：本协议一式八份，自双方签字盖章之日起生效，至项目完成之日起终止；若合作申请未获资助，本协议自动废止。

2211209643945

八、合同条款

第一条　甲方与乙方根据《中国人民共和国民法典》及国家有关法规和规定，为顺利完成2022年大数据智慧种养平台关键技术研究及其应用示范专项项目（文件编号：云科函〔2022〕88号）经协商一致，特订立本合同，作为甲乙双方在项目实施管理过程中共同遵守的依据。

第二条　甲方的权利义务：1.按合同书规定进行经费核拨的有关工作协调。2. 根据甲方需要，在不影响乙方工作的前提下，定期或不定期对乙方项目的实施情况和经费使用情况进行检查或抽查。

第三条　乙方的权利义务：1.确保落实自筹经费及有关保障条件。2. 按合同书规定，对甲方核拨的经费实行专款专用，单独列账，并随时配合甲方进行监督检查。3. 使用财政资金采购设备、原材料等，按照有关规定，符合招标条件的须进行招标。4. 项目实施完成或实施到一定程度，须提出验收或终止结题的申请，并按甲方要求做好项目结题工作。5. 在每年1月向甲方如实提交上年度工作情况报告，报告内容包含上年度项目进展情况、经费决算和取得的效果等。6. 按照国家、省和市、县（区）有关规定，提交科技报告或其他材料。

第四条　在履行本合同的过程中，如遇到1财政计划改变等不可抗力情况，甲方对所核拨经费的数量和时间可进行相应变更。

第五条　在履行本合同过程中，如项目完成的进度加快或延缓，经双方协商，可对合同中经费年度下达计划、项目进度和阶段目标进行相应变更。

第六条　在履行本合同的过程中，当事人一方发现可能导致项目整体或部分失败的情形时，应及时 通知另一方，并采取适当措施减少损失，没有及时通知并采取适当措施，致使损失扩大的，应当就扩大 的损失承担责任。

第七条　在履行本合同的过程中，因出现在现有水平下无法克服的技术困难，致使项目失败或部分失败造成损失，1. 乙方应及时通知甲方，提供相关证据并予以说明。2. 甲方以已核拔 的经费为最高限承担部分责任。

第八条　乙方违反约定造成项目工作停滞、延误或失败,未能通过验收,应承担违约责任。

第九条　本项目技术成果的归属、转让和实施技术成果所产生的经济利益的分享，除双方另有约 定外，按国家和有关法规执行。

第十条　属技术保密的项目当事人双方订立技术保密条款，作为合同正式内容的一部分。

第十一条　根据项目具体情况，经双方协商订立的附加条款作为本合同正式内容的一部分。

第十二条　甲方可根据具体情况决定乙方是否需要单位担保，若需要保证单位，应订立担保条款， 作为本合同正式内容一部分。

第十三条　本合同的争议应由双方本着协商一致的原则解决，当合同需要更改或解除时，双方应订 立变更条款或协议，仲裁和诉讼在甲方所在地进行。

第十四条　本合同一式四份，各份具有同等效力。甲方存二份，乙方存二份，本合同自签字之日起生效，有效期至项目验收后一年内。各方均应负合同的法律责任，不应受机构、人事变动而影响。

第十五条　违约责任：
违反本合同约定，违方应承担违约责任。
1. 违反本合同第三条第一项约定，乙方应当承担违约责任，承担方式和违约金额如下：
（1）合同解除。
（2）乙方退还甲方已核拨的经费，并自行承担由此引起的损失。
2. 违反本合同第八条约定，乙方应当承担违约责任，承担方式和违约金额如下：
（1）退还甲方已核拨的经费。
（2）按已核拔经费的20%支付违约金。

第十六条　保密条款：
1. 本合同保密内容范围为：
　无！

2. 本合同保密期限为：
　无！

3. 乙方应与可能知悉保密内容的人员签订技术秘密保护协议。
4. 各方应建立技术秘密保护制度：
5. 属技术保密的项目必须经市负责技术保密部门审查后，确定可否发表或用于国际合作和交流。

第十七条　保证条款（可由保证人和被保证人另行约定）
1. 当乙方不履行或不完全履行本合同，以及没有或没有完全承担违约责任时，保证人承担一般担保责任。
说明：
1. 本合同书中，凡是当事人约定无需填写的条款，应在该条款的空白处划（/）。
2. 委托代理人签订本合同书的，应出具合法、有效的委托书。

2211150964345

九、本合同签约各方

管理单位（甲方）：云浮市科学技术局　　（盖章）

单位地址：广东省云浮市吉祥中路25号市科技馆

法定代表人（或授权代表）：梁小鹏　　（签章）

联系人（经办人）姓名：苏英慧　　（签章）

E-mail：13826891813@qq.com

电话：13826891813

年　月　日

承担单位（乙方）：云浮市物联网研究院有限公司　　（盖章）

单位地址：云浮市新兴县新城镇沿江北路66号筠诚国际广场金融中心10楼

法定代表人（或法人代理）：钟淑琴　　（签章）

联系人（项目主管）姓名：黄晓宁　　（签章）

E-mail：47834322@qq.com

电话：0766-2929555

开户单位名称：云浮市物联网研究院有限公司

开户银行：广东南粤银行云浮新兴支行

开户帐号：660101230900000912

年　月　日

乙方主管部门（丙方）：新兴县科技局　　（盖章）

单位地址：广东省云浮市新兴县新城镇茅园新街茅园路46号四楼

法定代表人（或法人代理）：彭炳南　　（签章）

联系人（项目主管）姓名：新兴县科技局主管部门管理员　　（签章）

E-mail：752087444@qq.com

电话：18938121312

开户单位名称：新兴县科学技术局

开户银行：农村信用合作社

开户帐号：59380360120110034835

年　月　日

2211509643134 5

受理编号：211015176431530

项目编号：2021020302

文件编号：云科函〔2021〕84号

云浮市科技计划项目合同书

项目名称：	基于深度神经网络的猪只行为识别与分析研究		
计划类别：	专题三：智慧农业		
项目起止时间：	2021-05-01至2023-04-30		
管理单位（甲方）：	云浮市科学技术局		
承担单位（乙方）：	云浮市物联网研究院有限公司		
通讯地址：	云浮市新兴县新城镇沿江北路66号筠诚国际广场金融中心10楼		
邮政编码：	527400	单位电话：	0766-2929555
项目负责人：	毛宜军	联系电话：	13822275121
项目联系人：	黄晓宁	联系电话：	0766-2929555
乙方主管部门（丙方）：	新兴县科技局		

云浮市科学技术局

二零二一年制

一、项目实施内容

主要研究内容：本研究拟对猪的动作进行建模，尤其对猪的步态等进行建模和识别，并将步态识别与疾病防控结合起来，作为猪的疾病预测和预警的主要依据。对猪的步态动作建模和识别主要有两种范式：步态验证和步态鉴别。前者主要研究参考样本和测试样本是否代表同一个对象（猪），后者研究在一群候选样本中，找出与参考样本最接近的样本。由于猪的动作步态识别往往处于非配合场景中，本项目将关注非配合场景序列化信号等问题，以求解决在非配合、非直线行走等自然场景下步态动作识别准确率较低的问题。

拟解决的关键技术问题：1.由于行走和拥挤环境导致的部分遮挡的识别；设计一种深度耦合神经网络，该网络由两个相互卷积的神经网络构成，主要有以下4部分：双层卷积神经网络、分类损失函数和验证损失函数。该网络受限采用步态能量图作为输入，如图2所示，双层卷积神经网络共享参数权值，每个卷积神经网络分别包含三个基本操作单元，即卷积层、池化层和正则层。这样，由输入到卷积神经网络层的步态动作特征向量则分别由两条通路输入到验证损失层。然后再根据输入样本对的标签来优化网络各层参数。第四个卷积单元对前面两个卷积单元提取到的步态动作特征图进行差值的融合，生成的差值向量通过一个全连接层后输出到分类损失层中，最后基于类别标签和验证损失函数联合训练双层耦合神经网络。2.由于行走和摄像头移动导致的多角度变化的挖掘；本研究方案的核心是采用三层卷积神经网络来提取猪只的步态动作特征。本方案首先采用三联体动作步态样本查询、正样本和负样本作为卷积神经网络的输入，每一个神经网络都的第一个单元Conv1都包含了16个尺寸为4*4的过滤核，扩展步长为1，过滤后可以获得16个不同的特征矩阵。输出的特征矩阵输入到局部相应归一化层和池化层，池化核步长为2，尺寸是2*2。最后可以得到16个动作步态特征图。3.低质量猪只动作步态的时序特征深度挖掘；对猪只的动作步态识别和分类不仅仅要挖掘空间特征信息，还需要充分利用时序信息才能更有效地提高步态动作识别和分类的准确性。本研究拟通过增加长短期记忆单元的方式对研究方案二中的三联体网络进行升级改造，使之对时序信息的挖掘能更有效，从而提升猪只动作步态识别的总体准确率。

创新点：提出将深度神经网络应用到具体的畜牧物联网场景，提出通过视频动作非接触式识别来对猪只的动作和身体状态进行预测预警，并形成实际场景的应用示范。

二、项目考核指标

（一）项目完成后提供的研究开发成果及形式（须明确产品、专利、版权、标准等成果的类型及数量）

成果形式		成果数量	成果形式		成果数量
发明专利	申请	2	引进人才（人）		0
	授权	0	培养人才（人）		3
实用新型专利	申请	0	技术标准制定	牵头（个）	0
	授权	0		参与（个）	0
科技人才奖励（人）		0	科技报告（篇）		0
外观专利	申请	0	软件著作权（项）		3
	授权	0	论文论著（篇）		0
获国家级奖项（项）		0	被收录论文数（篇）	SCI	0
获省级奖项（项）		0		IE	0
新产品（或新材料、新装备、新品种（系））		1		CA	0
国外专利	PCT受理	0	新工艺（或新方法、新模式、新技术）		0
	授权	0	新服务（项）		0
服务企业数量（家）		技术服务数量（项）			0
		服务企业数量（家）			0

科技金融项目必填	开展培训宣讲活动场次（次）	0
	服务企业数量（家）	0
	帮助企业融资（万元）	0
	引进专业机构（家）	0
院士工作站项目必填	引进院士及其团队科技成果转化数量	0
	院士开展的战略咨询和技术指导次数	0
	院士年进站次数	0
	院士及院士团队年进站时间	
软科学项目必填	决策咨询报告（篇）（至少1篇）	0
	研究总报告（篇）	0
	研究中期报告（篇）	0
	研究分报告（篇）	0
	调研报告（篇）	0
	专著（篇）[须注明“**市软科学研究计划项目（项目编号：）资助”]	
	培养人才（人）	0
	获国家级奖项（项）	0
	获省级奖项（项）	0
	其他（具体形式：用户填）	0

其他成果及形式说明：

本项目的研究在畜牧养殖业中是一项应用创新，成果可以有效敏捷地嵌入到已有的生产系统，或在本项目的研究基础上快捷扩展成为独立的系统平台。本项目的研究成果和知识产权将通过智慧农业示范县等项目进行有效扁平化宣传推介，项目最终形成猪只异常行为在线监测预警系统1套，并在3个养殖场示范应用。

（二）主要技术经济指标及社会效益

累计新增销售收入（万元）	200
累计新增利税（万元）	50

其他主要技术经济指标及社会效益说明：

本项目完成后，将会产生猪只异常行为在线监测预警系统1套，形成核心算法和专利2件以上。研究成果的产出和知识积累将给畜牧养殖业带来全新的管理方式和预警预测方法，这些方法体系可以有效减轻畜牧养殖业过程中的动物状态监控工作量，为畜牧智慧养殖提供必不可少的核心组件和方法。项目成果将在3个示范场应用。

项目负责人（签章）：无宜军　　2021 年 10 月 20 日

三、项目进度和阶段目标

（一）项目起止时间：	2021-05-01 至 2023-04-30
（二）项目实施进度及阶段主要目标	
开始日期--结束日期	主要工作内容（限 500 字）
2021-05-01 至 2021-10-31	主要研究内容： 开展项目文献和数据整理，撰写基于深度神经网络的猪只行为识别与分析项目实施报告，同时对耦合网络进行深入研究，并实现该深度网络在猪的动作步态识别原型。 预期研究成果： （1）深度网络在猪只动作步态识别实施方案； （2）申请1项软件著作权登记；
2021-11-01 至 2022-04-30	主要研究内容： 对三联体网络进行深入研究，对多角度视频切换场景进行研究，并实现该深度网络在猪的动作步态识别原型。 预期研究成果： （1）申请专利1 项； （2）申请1项软件著作权登记；
2022-05-01 至 2022-10-31	主要研究内容： 对基于长短期记忆单元的递归神经网络进行深入研究，对不同的时序动作序列进行深入研究，并实现该深度网络在猪的动作步态识别原型。 预期研究成果： （1）申请1项软件著作权登记 （2）申请专利1项；
2022-11-01 至 2023-04-30	主要研究内容： 综合前面研究，理清相关方案中的各种细节问题，将各种深度学习模型进行有机整合，实现高性能、低消耗和高准确度的猪只动作步态分析挖掘原型系统。 预期研究成果 （1）项目累计获得3项软件著作权登记； （2）项目累计申请专利2项； （3）项目累计培养研究生3名； （4）获得猪只异常行为在线监测预警系统一套，在3个养殖场示范应用； （5）项目结题报告。

四、承担、参与单位工作分工及经费分配情况

承担/参与单位名称（盖章）	工作分工	总经费分摊（万元）	市科技局经费分配（万元）
云浮市物联网研究院有限公司	提供技术开发物联网数据和应用场景数据，负责技术数据模型研究，提供软硬件实验环境和业务专家组等协调、组织和沟通。	135	35
华南农业大学	负责数据挖掘、机器学习和人工智能技术研究，完成畜牧养殖数据到智能决策系统设计和原型实现。	7.5	7.5
暨南大学	负责大数据实时性处理研究，实现高性能技术框架和全流程数据安全技术架构研究。	7.5	7.5

五、项目总经费及市科技局经费预算

（一）市科技局经费下达总额：（大写） 伍拾万元整；（小写）：50 万元

（二）市科技局经费拨付方式： 一次性拨款

（三）市科技局经费经费年度下达计划： 第一期50万元

（四）总经费及市科技局经费开支预算计划：

经费筹集情况：（单位：万元）

总投入经费：150万元

	市科技局经费	自筹资金				合计
		自有资金	贷款	地方政府投入	其它	
已投入经费	0	0	0	0	0	0
新增经费	50	100	0	0	0	150

其它政府部门、境外资金及其他资金投入情况说明：

无

（二）新增经费预算（单位：万元）

	新增经费总额		其中：市科技局经费	
支出经费	经费额	用途说明	经费额	用途说明
基建费	0			
1、直接费用	147	主要包括设备费、材料费、燃料动力费、差旅费、会议费、国际合作与交流费、人员费、劳务费等	47	主要包括设备费、材料费、燃料动力费、差旅费、会议费、国际合作与交流费、人员费、劳务费等

(1) 设备费	28	购置本项目研发所需的设备高性能GPU服务器、数据节点服务器等	28	购置本项目研发所需的设备高性能GPU服务器、数据节点服务器等
(2) 材料费	40	项目相关研发，成果应用示范物料、监控摄像头、交换机、路由器等费用	4.5	项目相关研发，成果应用示范物料、监控摄像头、交换机、路由器等费用
(3) 测试化验加工外协费	0		0	
(4) 燃料动力费	10	研发过程中研发设备所发生的水电动力能源费，示范应用基地实验所需水、电等,由申报单位自筹	0	
(5) 差旅费/会议费/国际合作与交流费	3	相关差旅、会议及交流费用	3	相关差旅、会议及交流费用
(6) 出版/文献/信息传播/知识产权事务费	5	项目相关出版费、资料费、专利费、通信费、专用软件购买费、文献检索费及其他知识产权费用	5	项目相关出版费、资料费、专利费、通信费、专用软件购买费、文献检索费及其他知识产权费用
(7) 劳务费	9	项目参与人员劳务费	4.5	项目参与人员劳务费

（8）人员费	50	参与项目人员工资费用	0	
（9）专家咨询费	2	项目研发过程中的技术难点咨询	2	项目研发过程中的技术难点咨询
（10）直接费其他支出	0		0	
（11）科技金融服务体系其他费用	0		0	
①信用评级补贴	0		0	
②大赛场租	0		0	
③特派员奖励与补贴	0		0	
2、间接费用	3	项目间接费用	3	项目间接费用
（1）间接成本	0		0	
（2）管理成本	2	项目管理费用支出	2	项目管理费用支出
（3）绩效支出	1	项目激励及绩效支出	1	项目激励及绩效支出
合计	150		50	

六、人员信息

项目负责人：

姓名	性别	年龄	职务	职称	学历	在项目中承担的任务	所在单位	签名
毛宜军	男	42	企业科技特派员	讲师	博士研究生	项目全面技术和管理工作，负责项目统筹实施，把控项目研发方向及进度。	云浮市物联网研究院有限公司	毛宜军

主要研究开发人员：

姓名	性别	年龄	职务	职称	学历	在项目中承担的任务	所在单位	签名
郭杰	男	46	工程中心主任	经济师	硕士研究生	项目管理，软硬件实验环境和专家组等协调、组织和沟通。	云浮市物联网研究院有限公司	郭杰
周皓恩	男	36	研发部总监	未取得	硕士研究生	技术数据模型研究	云浮市物联网研究院有限公司	周皓恩
黄晓宁	男	39	科研管理	助理工程师	本科	基于深度神经网络的猪只行为识别与分析系统应用推广	云浮市物联网研究院有限公司	黄晓宁
叶志峰	男	26	研发人员	未取得	本科	物联网数据和应用场景数据开发	云浮市物联网研究院有限公司	叶志峰
梁梓琪	男	26	研发人员	未取得	本科	物联网数据和应用场景数据开发	云浮市物联网研究院有限公司	梁梓琪
伍晓仪	女	27	科研管理员	未取得	本科	协调项目实施，协助研发过程中各项工作。	云浮市物联网研究院有限公司	伍晓仪
古万荣	男	39	教师	讲师	博士研究生	深度学习模型研究与实现，猪只行为视频特征提取与建模	华南农业大学	古万荣

刘心	女	44	教师	副教授	博士研究生	物联网数据和应用场景数据开发	华南农业大学	刘心
黄栋	男	34	教师	副教授	博士研究生	超大规模数据处理	华南农业大学	黄栋
王国华	男	33	教师	讲师	博士研究生	视频动作识别与挖掘	华南农业大学	王国华
崔金荣	女	37	教师	讲师	博士研究生	视频动作识别与挖掘	华南农业大学	崔金荣
张继连	男	44	教师	副教授	博士研究生	物联网大数据管理方案设计与建模	暨南大学	张继连
于欣宇	女	25	研究生	未取得	本科	数据采集算法设计与实现	暨南大学	于欣宇
唐蓉	女	24	研究生	未取得	本科	大数据安全存储方案研究	暨南大学	唐蓉
王宋凯	男	23	研究生	未取得	本科	分布式大数据存储方案实现	暨南大学	王宋凯
王钰	女	23	研究生	未取得	本科	大数据高效检索系统设计	暨南大学	王钰
杨良勇	男	25	研究生	未取得	本科	大数据高效检索系统实现	暨南大学	杨良勇
陈志彬	男	23	研究生	未取得	本科	大数据知识图谱研究与实现	暨南大学	陈志彬

七、承担、参与单位合作协议（须与申报书中合作协议或意向书相一致）

联合申报项目协议书

项目申报单位（甲方）：云浮市物联网研究院有限公司
项目参与单位（乙方）：华南农业大学
项目参与单位（丙方）：暨南大学

甲乙丙各方就共同申报云浮市2021年省科技专项资金（“大专项+任务清单”）中的“基于深度神经网络的猪只行为识别与分析研究”项目的相关事宜，经友好协商，达成如下合作协议：

第一条：项目研究工作详细分工：
甲方（主持方）：提供技术开发物联网数据和应用场景数据，负责技术数据模型研究，提供软硬件实验环境和业务专家组等协调、组织和沟通。
乙方（参与方）：负责数据挖掘、机器学习和人工智能技术研究，完成畜牧养殖数据到智能决策系统设计和原型实现。
丙方（参与方）：负责大数据实时性处理研究，实现高性能技术框架和全流程数据安全技术架构研究。

第二条：经费分配：
1、如果本申报项目获批立项，按各级政府下达的资助经费综合，甲方、乙方和丙方同意此经费分别按政府资助经费的甲方：70%;乙方：15%;丙方：15%进行分配。
2、甲方在收到云浮市科学技术局下达的资助经费后的一个月内将乙方和丙方所占经费分别支付给乙方和丙方指定账户。
乙方账户：华南农业大学
乙方账号：3602002609000310520
开户行：中国工商银行广州市五山支行

丙方账户：暨南大学
丙方账号：3602015819100000858
开户行：中国工商银行广州暨南大学支行

第三条：课题组成员：甲方单位项目负责人毛宜军，联系电话为13822275121;乙方单位项目负责人为古万荣，联系电话为13719398640,丙方单位项目负责人为张继连，联系电话为18278382172。

第四条：知识产权归属：项目实施过程中所产生的知识产权，优先执行任务下达单位的知识产权管理政策，在此前提下，各方独立完成的所有权归各自所有；双方共同完成的由双方共享，具体按照双方的贡献大小进行分配或双方另行商定。项目成果申报各级奖项，双方单位排名根据具体情况另行商定，人员排名原则上按贡献大小先后排名。

第五条：本协议一式八份，自双方签字盖章之日起生效，至项目完成之日起终止；若合作申请未获资助，本协议自动废止。

八、合同条款

第一条　甲方与乙方根据《中国人民共和国民法典》及国家有关法规和规定，为顺利完成2021年基于深度神经网络的猪只行为识别与分析研究专项项目（文件编号：云科函〔2021〕84号）经协商一致，特订立本合同，作为甲乙双方在项目实施管理过程中共同遵守的依据。

第二条　甲方的权利义务：1.按合同书规定进行经费核拨的有关工作协调。2. 根据甲方需要，在不影响乙方工作的前提下，定期或不定期对乙方项目的实施情况和经费使用情况进行检查或抽查。

第三条　乙方的权利义务：1.确保落实自筹经费及有关保障条件。2. 按合同书规定，对甲方核拨的经费实行专款专用，单独列账，并随时配合甲方进行监督检查。3. 使用财政资金采购设备、原材料等，按照有关规定，符合招标条件的须进行招标。4. 项目实施完成或实施到一定程度，须提出验收或终止结题的申请，并按甲方要求做好项目结题工作。5. 在每年1月向甲方如实提交上年度工作情况报告，报告内容包含上年度项目进展情况、经费决算和取得的效果等。6. 按照国家、省和市、县（区）有关规定，提交科技报告或其他材料。

第四条　在履行本合同的过程中，如遇到1财政计划改变等不可抗力情况，甲方对所核拨经费的数量和时间可进行相应变更。

第五条　在履行本合同过程中，如项目完成的进度加快或延缓，经双方协商，可对合同中经费年度下达计划、项目进度和阶段目标进行相应变更。

第六条　在履行本合同的过程中，当事人一方发现可能导致项目整体或部分失败的情形时，应及时 通知另一方，并采取适当措施减少损失，没有及时通知并采取适当措施，致使损失扩大的，应当就扩大 的损失承担责任。

第七条　在履行本合同的过程中，因出现在现有水平下无法克服的技术困难，致使项目失败或部分失败造成损失，1. 乙方应及时通知甲方，提供相关证据并予以说明。2. 甲方以已核拨 的经费为最高限承担部分责任。

第八条　乙方违反约定造成项目工作停滞、延误或失败,未能通过验收,应承担违约责任。

第九条　本项目技术成果的归属、转让和实施技术成果所产生的经济利益的分享，除双方另有约 定外，按国家和有关法规执行。

第十条　属技术保密的项目当事人双方订立技术保密条款，作为合同正式内容的一部分。

第十一条　根据项目具体情况，经双方协商订立的附加条款作为本合同正式内容的一部分。

第十二条　甲方可根据具体情况决定乙方是否需要单位担保，若需要保证单位，应订立担保条款，作为本合同正式内容一部分。

第十三条　本合同的争议应由双方本着协商一致的原则解决，当合同需要更改或解除时，双方应订 立变更条款或协议，仲裁和诉讼在甲方所在地进行。

第十四条　本合同一式四份，各份具有同等效力。甲方存二份，乙方存二份，本合同自签字之日起生效，有效期至项目验收后一年内。各方均应负合同的法律责任，不应受机构、人事变动而影响。

第十五条　违约责任：
违反本合同约定，违方应承担违约责任。
1. 违反本合同第三条第一项约定，乙方应当承担违约责任，承担方式和违约金额如下：
（1）合同解除。
（2）乙方退还甲方已核拨的经费，并自行承担由此引起的损失。
2. 违反本合同第八条约定，乙方应当承担违约责任，承担方式和违约金额如下：
（1）退还甲方已核拨的经费。
（2）按已核拔经费的20%支付违约金。

第十六条　保密条款：
1. 本合同保密内容范围为：
　无

2. 本合同保密期限为：
　无

3. 乙方应与可能知悉保密内容的人员签订技术秘密保护协议。
4. 各方应建立技术秘密保护制度：
5. 属技术保密的项目必须经市负责技术保密部门审查后，确定可否发表或用于国际合作和交流。

第十七条　保证条款（可由保证人和被保证人另行约定）
1. 当乙方不履行或不完全履行本合同，以及没有或没有完全承担违约责任时，保证人承担一般担保责任。
说明：
1. 本合同书中，凡是当事人约定无需填写的条款，应在该条款的空白处划（/）。
2. 委托代理人签订本合同书的，应出具合法、有效的委托书。

九、本合同签约各方

管理单位（甲方）：云浮市科学技术局　　（盖章）

单位地址：广东省云浮市吉祥中路25号市科技馆

法定代表人（或授权代表）：梁玉莉　　（签章）

联系人（经办人）姓名：苏英慧　　（签章）苏英慧

E-mail：13826891813@qq.com

电话：13826891813

2021年10月29日

承担单位（乙方）：云浮市物联网研究院有限公司　　（盖章）

单位地址：云浮市新兴县新城镇沿江北路66号筠诚国际广场金融中心10楼

法定代表人（或法人代理）：钟淑琴　　（签章）

联系人（项目主管）姓名：黄晓宁　　（签章）

E-mail：47834322@qq.com

电话：0766-2929555

开户单位名称：云浮市物联网研究院有限公司

开户银行：广东南粤银行云浮新兴支行

开户帐号：660101230900000912

2021 年 10月20日

乙方主管部门（丙方）：新兴县科技局　　（盖章）

单位地址：广东省云浮市新兴县新城镇茶园新街茶园路46号四楼

法定代表人（或法人代理）：彭炳南　　（签章）

联系人（项目主管）姓名：新兴县科技局主管部门管理员　　（签章）

E-mail：752087444@qq.com

电话：15752087444

开户单位名称：新兴县科学技术局

开户银行：农村信用合作社

开户帐号：59380360120110034835

2021 年 10月20日

HXKJHT2023283

衢州柑橘大数据分析建模平台
建设合作合同书

甲方：原动力（广州）技术有限公司　　（下简称甲方）

乙方：华南农业大学　　（下简称乙方）

一、总则

依据《中华人民共和国民法典》的有关规定，为明确合同甲方和乙方在本项目中的权利和义务，保证本项目的顺利开展，根据国家法律、法规，甲、乙双方遵循平等、自愿、公平和诚实信用原则，经友好协商，特订立本合同。

二、释义

本合同中使用的下列词语，除非合同中明确另有所指，具有如下含义：

2.1“系统”、“信息系统”等均指本合同所规定的、甲方拟投资建设、乙方提供的相关产品和服务的信息技术系统。还特指本合同涉及的软件、相关技术资料、文档等的统称。

2.2 “合同”是指甲乙双方订立的协议，包括合同条款及其所有附件。

2.3 “合同价格”是指在本合同项下由甲方支付给乙方的全部报酬。

2.4 “软件”除另有指明外，指本合同履行期内所集成和提供的当前和将来的软件版本，包括乙方为履行本合同所集成和提供的软件版本和相关的文件。

2.5 “服务”指的是用于调研、设计、装配、安装、技术指导、调试、试运行、工程验收、性能保证、维护保修、人员培训及合同项下乙方应承担的其他有关服务。

2.6 “技术文件”或“文档”指与合同系统运行、维护和检验有关的技术参数、图纸、设计、手册、专有信息与合同系统计算、系统的安装、调测、运行、维护和检验相关的文件。

2.7 “源代码”指用于该软件的源代码。其必须可为熟练的程序员理解和使用，以可打印及被机器阅读或具备其他合理而必要的形式存在，包括对该软件的评估、测试或其它技术文件。

2.8 “培训”“技术培训”指按合同的规定，由乙方/乙方代表或者乙方指定的第三方独立培训机构通过理论讲解和现场指导等方式将相关知识传授给甲方受训（指定）人员的行为。

2.9 质量保证期：是指系统上线并经甲方初步验收评审后三年。

2.10 “甲方现场”或“现场”指对合同系统进行安装和运行的场所。

2.11 “商业秘密”指甲、乙方各自所拥有的，不为公众所知的管理信息、方式方法、客户名单、商业数据、产品信息、技术诀窍、源代码、计算机文档等，或由甲、乙方在履行本合同过程中明确指明为商业秘密的、法律所认可的任何信息。

三、合同标的

3.1 本合同开发的内容为“衢州柑橘大数据分析建模平台”。乙方提供自有版权或授权的软件及二次开发工作为甲方建设此系统。乙方保证其提供的产品及服务能全部满足甲方需求。

3.2 本合同项目实施：本项目要求按照技术需求说明书完成项目的全面上线工作。

四、双方权利和义务

4.1 甲方权利和义务

4.1.1 甲方保留在系统开发过程中根据自身业务需求，对尚未执行的合同项进行业务需求调整变更的权力。

4.1.2 甲方有权建立相应的组织机构，落实项目各阶段工作内容，协调乙方有关人员，参与开展实施、培训、上线工作。

4.1.3 甲方有权根据本合同的规定和项目需要，向乙方提供有关资料，并指派相关人员配合乙方调查了解有关情况，需安排专业技术人员专职配合乙方工作，从合同签订后进入。

4.1.4 甲方有权按进度要求组织各里程碑成果的评审工作。

4.1.5 甲方应按本合同规定支付合同费用。

4.1.6 甲方应遵守本合同所规定的知识产权与保密方面的义务。

4.2 乙方权利和义务

4.2.1 如甲方购买实施服务后，乙方需严格根据合同文件及甲方要求形成项目实施详细计划，并与甲方沟通并经甲方书面同意后予以实施。

4.2.2 按照附件约定的时间完成合同系统的需求调研、开发、实施、培训工作，并提供及时的免费系统维护服务。

4.2.3 乙方有义务对甲方现有的技术设备及软件进行充分了解，并结合甲方的其他系统，充分利用甲方已有设备和相关软件开展工作。除甲方书面明确表示需依照本合同由乙方升级的和不予保留的功能之外，乙方不得由于集成了本合同所规定的系统后，而破坏其原有功能。

4.2.4 乙方有义务协助甲方提出适合甲方系统的软硬件配置方案。

4.2.5 乙方负责对甲方技术人员和业务部门使用人员等进行产品的现场培训。

4.2.6 遵守本合同规定的知识产权与保密方面的义务。

五、合同范围

“衢州柑橘大数据分析建模平台”中的信息技术服务：“人工智能训练平台”的开发、实施、运维服务，并配合甲方进行产品化的规划、设计与开发工作。详细需求参照业主单位签字确认的技术文件。

六、价格及支付方式

本合同的总金额为人民币元￥29.6万元（大写：贰拾玖万陆千元整）。合同各方承担各自的税务责任，甲方不因签署本合同而导致连带承担乙方的税务责任，如果因此而导致甲方的损失，乙方应当赔偿。本合同价格包含3年质保期内运行维护及技术支持服务。

付款方式为分期付款，具体如下：

第一笔：项目合同生效且乙方技术实施方案经甲方确认后10个工作日内，甲方需支付合同总价的 20%。（该项费用已由甲方支付给上一任供应商 上海伫鹏信息科技中心）

第二笔：完成系统全面上线并经甲方初验合格后30个工作日内，甲方需向乙方支付合同总价的 60%。

第三笔：项目试运行一年维护后甲方需向乙方支付合同总价的 10%。

第四笔：项目试运行三年维护且完成终验后甲方需向乙方支付合同总价的10%。

3年期质保及技术支持服务满后，若甲方需乙方继续提供技术支持服务和延续质保，需向乙方支付合同金额 6 %/年技术支持服务和延续质保费用。

乙方按照下列信息向甲方开具增值税普通发票：

开户银行：招商银行股份有限公司广州淘金支行

户名：原动力（广州）技术有限公司

帐号：120917176810601

地址：广州市海珠区庭园路3号2206房

电话：18922430967

纳税人识别号：91440101MA5D4PHKX9

甲方按照下列信息向乙方支付合同款：

乙方法定名称： 华南农业大学

统一社会信用代码： 124400004554165634

公司银行账号：3602002609000310520

开户行： 广州工行五山支行

电话： 13719398640

联系人（项目负责人）： 古万荣

合同各方的账户名称、开户银行、银行账号以本合同提供的上述信息为准。合同各方如未正确依照上述规定支付合同的相关款项或者开具发票，应及时沟通协商解决。

七、项目任务、要求及范围

7.1 本合同项下项目的任务、要求和范围见技术规范书。

7.2 组成本合同的文件包括：

（1） 本合同条款；

合同履行中，双方有关项目的洽商、变更等书面协议或文件、安全生产协议书、履约保函等均视为本合同的组成部分。上述文件是相互说明、相互解释的，如出现不一致时，上述的排列顺序就是合同的优先解释和执行顺序。

八、知识产权

8.1 乙方应保证，甲方使用乙方提供的产品及服务的全部及任何一部分时，免于遭受任何第三方因此而提出的侵权诉讼及索赔。如果任何第三方向甲方提出侵权索赔，乙方应负责与之交涉，并承担由此造成的一切损失。

8.2 本合同项目中乙方为甲方专有应用需求而定制开发的甲方专用软件模块，知识产权归甲方所有。

8.3 本合同进行中，包括但不限于二次开发产生的软件和文件版权（乙方为甲方专有应用需求而定制开发的甲方专用软件模块除外）由双方共同所有。非经对方书面同意，一方不得将该软件及相关资料以任何形式赠送、出售给第三方，但双方的子公司、分公司或下属机构除外。

8.4 本合同产品甲乙双方均有权进行后续开发，互不干涉，亦不从另一方的独立经营中获取利益。乙方向甲方提供此系统方面的文档及技术系统实现，保证甲方在此平台上继续开发。

九、保密条款

9.1 合同各方对在签订和履行本合同过程中从任一方获知的对方的技术和商业秘密，无论在本合同期限内还是合同终止后，均应共同遵守国家有关版权、

专利、商标等知识产权方面的法律规定及各部委颁布的保密规定中的相关条文。未经一方事先书面授权，任何一方不得以任何方式向任何其他组织或个人泄露、转让、许可使用、交换、赠与或与任何其他组织或个人共同使用或不正当使用。违反本条规定，给对方造成损失的，违约方应向对方承担相应的民事赔偿责任。

9.2 本合同所述保密信息指：本合同一方（“信息披露方”）向本合同另一方（“信息接受方”）按照本合同（或就本合同）所提供的各类技术和商业信息（包括甲方的业务数据、商业运作模式等）、技术规格、图纸、文件、信息及专有技术、客户信息。保密信息被视为信息披露方的专有财产。

9.3 当保密信息出现下述情况时，本章对保密信息的限制不再适用：

并非信息接受方的过错而已经进入公知领域的；

已通过信息接受方的有关记录证明是由信息接受方独立开发的；

法律法规或监管部门要求信息接受方披露的，但信息接受方应在合理的时间内提前通知信息披露方，使其得以采取其认为必要的保护措施。

9.4 因执行本合同，各方均有权向其关联公司人员及职员透露或使其接触从披露方获取的保密信息的全部或其中的任何部分，范围是这些职员应是项目组成员、项目管理人员、项目成果直接应用人员、项目领导人员及为执行本合同必须知悉该保密信息的人员，前提是接受保密信息的一方向职员透露或使其接触前已经从其职员获得了必要的保密和不透露的承诺。任何一方的关联公司人员及该方职员违背上述承诺，向第三方披露保密信息，或依据该等保密信息向第三方作出任何建议，都被视为该方违反本合同规定。

9.5 本章节规定的保密义务在本合同终止或期满后（三年）继续有效。

9.6 任何一方违反本章约定均被视为违约。违约方应承担违约责任。

十、合同的生效、变更和终止

10.1 本合同自双方法定代表人或授权代表签字或盖章完成后生效。未尽事宜由双方友好协商一致后签订补充协议解决。本合同在双方完成合同规定的责任义务后终止。

10.2 本合同一式四份，其中甲方二份，乙方二份。本合同所列的附件、及项目实施过程中所出具的书面文件，经双方加盖公章后，为本合同的组成部分，并具有与合同正文同样的法律效力；其它任何口头或未包含在本合同内的、或未依据本合同制定的书面文件，均不对双方发生拘束力。如果本合同附件与合同正文不一致，合同正文效力优先。

10.3 本合同一经签署，未经双方同意，任何一方不得更改本合同。

10.4 在合同执行过程中，任何一方可以书面形式向对方提出对本合同的变更请求，另一方应在收到书面请求之日起一周内与提出方进行协商。双方协商达成一致之前，不影响原合同的履行，达成一致后可以用书面补充合同的方式对原合同进行变更。

10.5 没有甲方的事先书面同意，乙方不得将本合同项下的任何权利和义务全部或部分转让给第三方。

10.6 双方同意解除合同或根据本合同及相关法律法规应终止并解除合同时，甲乙双方应签订书面文件，处理善后事宜。

10.7 如发生以下情况，任何一方有权终止合同，但须以书面方式及时通知对方：

（1）一方进入破产、解散、被依法关闭撤销或已进入清算阶段；

（2）出现了合同规定的或法定解除事由。

除本合同和法律法规另有规定外，任何一方发生上述情况，将被视作违约，另一方有权依照本合同的规定，追究该方的违约责任。

10.8 合同到期后，合同项下任何尚未了结的权利和义务不受合同到期的影响，债务人仍有义务向债权人继续履行其义务。

十一、违约责任

11.1 工期：合同生效后 60 日历天。如果因乙方原因不能按期完成任务，则每延期1天，乙方应当向甲方支付本合同合同总金额的0.1%作为违约金，且乙方同意甲方有权从其应向乙方支付的相关费用中直接扣除上述违约金。如果

因甲方原因不能按合同约定向乙方支付合同款项，则每延期1天，甲方应当向乙方支付本合同合同总金额的0.1%作为违约金，且甲方同意在应向乙方支付的相关费用中直接累加上述违约金。

11.2 如发生违约事件，守约方要求违约方支付违约金时，应以书面方式通知违约方，内容包括违约事件、违约金、支付时间和方式等。违约方在收到上述通知后，应于5个工作日内答复对方，并支付违约金。若违约方在10个工作日内不予答复，视为对方默认违约事实。

如双方不能就此达成一致意见，将按照本合同所规定的争议解决条款解决双方的纠纷。

11.3 甲、乙双方承担违约责任后，仍需继续履行合同所规定的所有责任及义务。

11.4 乙方保证所提供的软件及伴随服务，应达到本合同所有组成文件约定的标准。如果因乙方的原因，软件及伴随服务不能达到合同所有组成文件约定的标准，乙方应立即整改，如若仍不能达到合同所有组成文件约定的标准，甲方有权寻求其它供应商完成相应整改，但产生的费用由乙方承担。乙方承担本项费用，不超过甲方向乙方实际支付的合同款项的10%。

十二、争议解决

甲乙双方因本合同引起的或与本合同有关的任何争议，应首先争取通过友好协商解决，如果在一方向另一方发出要求友好协商的书面通知后30日内，争议没有通过友好协商解决，任何一方均可向合同签订地法院起诉。

十三、通知

13.1 为享有本合同所规定的权利及履行本合同所规定的义务或有关违约交涉而需通知另一方时，通知方须采取书面形式，以挂号信件或以专人快递送达方式送达被通知方。挂号信寄出7工作日后，视为已经送达；专人送达的，被送达人签署后，视为已经送达。

13.2 通知地址为本合同签订双方注册地址。如一方欲改变通知地址，则应提前以书面方式通知另一方。

十四、不可抗力

14.1 由于地震、台风、水灾、火灾、战争以及其他不能预见并对其发生和后果不能预防或避免的不可抗力，直接影响本合同的履行或者不能按照合同的约定履行时，遇有上述不可抗力的一方，应立即书面通知对方，并在十五天之内，提供上述不可抗力的详细情况及合同不能履行，或者部分不能履行，或者需要延期履行的理由和有效的证明文件。此项证明文件应由不可抗力发生地的公证机关出具。按其对履行合同影响的程度，由双方协商决定是否解除合同，或者部分免除履行合同的义务，或者延期履行合同。

14.2 受到不可抗力影响的一方，应尽可能地采取合理的行为减轻不可抗力对履行本合同的影响。

十五、验收标准

最终开发需求范围及验收标准根据《衢州市柯城区农业大数据平台建设项目》招标文件中功能需求内容为准。

甲方：原动力（广州）技术有限公司　　乙方：华南农业大学

（盖章）　　（盖章）

负责人：　　负责人：

签订日期（生效日期）：2023年12月1日

HXKJHT20232087

智慧两化党支部大数据平台合同

甲方：广州市联展贸易有限公司

乙方：华南农业大学

签订日期：2023 年 8 月 15 日

甲乙双方本着平等互利、协商一致的原则签订本合同，以供双方信守执行。

一、 本合同的构成

附录是本合同不可分割的组成部分，与本合同具有同等效力。

双方均知悉合同附录的内容。乙方同意按照本合同和附件的规定提供软件，甲方同意接受该等文件的约束。

二、 合同价款及支付方式

2.1 本合同总金额为 50,000.00 元，大写：伍万元整。

2.2 本合同项下款项的币种为 人民币 。

2.3 付款方式：

第一期付款：合同协议签订后，甲方向乙方支付至合同核定金额的 30%，即：【15,000.00】元，大写：【壹万伍仟元整】，乙方收到款项后，安排约定范围内相关开发并配合甲方测试；

第二期付款：平台通过甲方测试后，乙方把平台部署至甲方指定云平台，部署完成验收后，甲方向乙方支付至合同核定金额的 60%，即：【35,000.00】元，大写：【叁万元整】；

第三期付款：平台运行 1 月内没有问题，7 日内甲方支付剩余合同金额给乙方，即：【5,000.00】元，大写：【伍仟元整】

2.4 项目时间：

从签完合同首付款到账时间起，合同生效且甲方 20 个工作日内完成。

2.5 甲方账号信息：

公司名称：广州市联展贸易有限公司

纳税人识别号：91440112593702234R

地址：广州市黄埔区荔联街东逸三街 18 号 301 房（仅限办公用途）

电话：13922382010

开户行：中国建设银行广州东环支行

账号：44050149120100001255

2.6 乙方账号信息：

户名（公司名称）： 华南农业大学

开户银行：广州工行五山支行

银行账号：3602 0026 0900 0310 520

2.7 差旅费用：

以上报价总价已经包含项目整体验收的一次一人差旅费用（注：截止到验收交付用户，如果乙方根据甲方要求出现场支持超过一次则从第二次开始乙方提供对应差旅票据由甲方根据费用发票或凭证实报实销）。

2.8 发票：

甲方每次付款前需要乙方提供对应金额的发票，税率：6% ，发票类型：增值税专用发票。

三、 软件的开发及交付

3.1 本合同期限分制作期限和维护期限。

制作期限自甲方提供完整的书面材料之后，项目实施总时间预计为 20 个工作日，其为程序开发时间；

维护时间是壹年，从乙方完成制作且甲方验收合格之日起，至年维护的期限结束。

3.2 软件的开发进度及开发流程以本合同“附录”为准，但因甲方变更软件需（要）求、延迟确认或验收等非乙方原因导致开发进度变更的，相应的开发进度予以顺延，甲方对此不持异议。

3.3 如用户需要本次合同项下内容的源代码，则乙方按要求提供，且不额外增加费用。

四、 安装及验收

4.1 验收标准以本合同的“附录”为验收标准；

4.2 乙方将平台部署到甲方指定的云平台，通知甲方进行验收确认。甲方如有异议，应在收到通知后三个工作日内向乙方提交书面文件。甲方逾期未向乙方递交书面文件的，视为本合同项下软件已通过甲方验收，符合本合同附件对平台的需（要）求，为合格产品。

五、 甲方的权利与义务

5.1. 甲方应当按约付款。如甲方违反付款义务的，乙方有权中（终）止开发进度，直至甲方完成付款义务，因开发进度中（终）止所造成的后果由甲方自行承担。

5.2. 甲方有权获知乙方的开发进度并提出修正意见，有义务对乙方的开发进度进行确认和验收。

5.3. 原则上，甲方不得对本合同附件所约定的软件开发需求（要求）进行修改。甲方确需变更的，由甲乙双方另行协商开发进度、费用等事宜。

六、 乙方的权利与义务

6.1. 乙方保证对本合同项下软件拥有完整的知识产权，不存在侵犯他人知识产权的情形，否则由此产生的责任全部由乙方承担。

6.2. 乙方有权要求甲方按约付款，如甲方违反付款约定，乙方有权视情形中（终）止开发，并追究甲方相应的违约责任。

6.3. 除非双方另有书面约定或法律有强制性规定，乙方仅按照本合同的约定承担责任。

七、 违约责任

7.1. 一方违反本合同（包括附件）的约定给另一方造成损失的，应当赔偿另一方的损失，赔偿金额以本合同总金额为限，但本合同另有约定的除外。

八、 争议解决

8.1. 由本合同引起的或与本合同有关的任何争议，双方应当友好协商解决；如协商不成，双方均可向广州仲裁院申请仲裁，仲裁败诉方承担胜诉方发生的全部费用，包括但不限于仲裁费、保全费、律师费、交通费、鉴定费、差旅费等。

九、 其他

9.1. 合同一式两份，甲乙双方各执壹份，具有同等法律效力，经双方签章后生效。

9.2. 本合同所列附件为本合同不可或缺的部分，附件内容与本合同主文具有同等法律效力。

9.3. 本合同指定乙方项目负责人是乙方单位华南农业大学古万荣。

甲方签章：

签章日期：2023.8.15

乙方签章：

签章日期：2023.8.15

附录

智慧两化党支部大数据平台技术要求及参数

智慧两化党支部大数据平台功能清单

产品	品牌/型号	技术参数/功能描述
一、平台		
党建学习平台	组织建设	1、查看并管理人员信息; 2、根据在职在册，入党及转正时间进行管理展示(如所属队伍，党籍状态之类)； 3、 录入人员更多详细信息，如荣誉，联系方式，用工方式等; 4、个人及支部信息维护; 5、支部历程及大事记管理。

		1、分配管理员及用户角色; 2、新增及管理各党委员会信息; 3、展示党组织数据及相应委员会联系人、组织书记，成立时间等信息; 4、个人及支部信息展示; 5、统计现委员会数量，党总支数，党支部数。
	理论学习	1、制订年度/年度及月度学习计划; 2、 学习待办提醒; 3、学习教育记录，党员学习笔记、体会记录等; 4、党支部每季度对党员督导检查及学习情况记录; 5、学习资源库管理; 6、学习积分管理。
	组织生活	1、会议通知管理; 2、三会一课管理; 3、会议签到及表决管理; 4、谈心谈话管理; 5、主题党日活动;
		6、年终工作考评管理; 7、统计分析管理。
	制度库	1、三会一课”制度、组织生活会制度、晋位升级制度、党风廉政建设责任制度库管理; 2、制度上墙管理。
	信息管理	1、每日一句管理及维护; 2、信息采编、管理及发布; 3、推荐及滚动信息管理及展示; 4、主题党日活动及主题教育活动; 5、阵地建设管理等。
	论坛与留言	1、可以发布问题，用户回复及评论; 2、论坛帖子管理; 3、留言发布与回复; 4、新内容提醒。
	台账管理	1、工作台账管理; 2、党员教育管理台账管理; 3、发展党员台账及其他台账管理
	管理中心	1、登录及注销、组织机构、用户管理、角色管理、群组管理、表单管理、流程管理、流程监控、系统菜单、消息管理、数据字典等。
	测试及上线	内部测试、优化及上线
二、售后服务		
售后服务		1、一年合同质保期内的免费平台维护，团队远程微信、QQ、电话支撑;

HXKJHT20241029

面向海量数据的 Mooze 商城推荐系统

开发合同

甲方：广州联展科技有限公司

乙方：华南农业大学

签订日期：2024 年 1 月 2 日

甲乙双方本着平等互利、协商一致的原则签订本合同，以供双方信守执行。

一、 本合同的构成

附录是本合同不可分割的组成部分，与本合同具有同等效力。

双方均知悉合同附录的内容。乙方同意按照本合同和附件的规定提供软件，甲方同意接受该等文件的约束。

二、 合同价款及支付方式

2.1 本合同总金额为 85,000.00 元，大写：捌万伍仟元整。

2.2 本合同项下款项的币种为 人民币 。

2.3 付款方式：

第一期付款：合同协议签订后，甲方向乙方支付 30,000.00 元，大写：【叁万元整】，乙方收到款项后，安排约定范围内相关开发并配合甲方测试；

第二期付款：平台通过甲方测试后，乙方把平台部署至甲方指定云平台，部署完成验收后，甲方向乙方支付 45,000.00 元，大写：【肆万伍仟元整】；

第三期付款：平台运行 1 月内没有问题，7 日内甲方支付剩余合同金额给乙方，即： 10,000.00 元，大写：【壹万元整】

2.4 项目时间：

从签完合同首付款到账时间起，合同生效且甲方 25 个工作日内完成（2024 年 2 月 9 号前部署上线；需求明确不再变更，如由于需求变更导致项目延期，工期顺延）。

2.5 乙方账号信息：

帐号户名：华南农业大学

开户银行：广州工行五山支行

银行账户：3602 0026 0900 0310 520

以上报价不含驻场或出差费用，如需驻场或出差，费用另议。

乙方项目负责人：古万荣，为乙方在编在岗员工。未经甲方书面同意，乙方不得更改项目负责人。

2.6 发票：

发票类型：增值税专用发票。

三、 软件的开发及交付

3.1 本合同期限分制作期限和维护期限。

制作期限自甲方提供完整的书面材料之后，项目实施总时间预计为25个工作日，其为程序开发时间；

维护时间是壹年，从乙方完成制作且甲方验收合格之日起，至年维护的期限结束。

3.2 软件的开发进度及开发流程以本合同“附录”为准，但因甲方变更软件需（要）求、延迟确认或验收等非乙方原因导致开发进度变更的，相应的开发进度予以顺延，甲方对此不持异议。

四、 安装及验收

4.1 验收标准以本合同的“附录”为验收标准；

4.2 乙方将平台部署到甲方指定的云平台，通知甲方进行验收确认。甲方如有异议，应在收到通知后三个工作日内向乙方提交书面文件。甲方逾期未向乙方递交书面文件的，视为本合同项下软件已通过甲方验收，符合本合同附件对平台的需（要）求，为合格产品。

五、 甲方的权利与义务

5.1. 甲方应当按约付款。如甲方违反付款义务的，乙方有权中（终）止开发进度，直至甲方完成付款义务，因开发进度中（终）止所造成的后果由甲方

自行承担。

5.2. 甲方有权获知乙方的开发进度并提出修正意见，有义务对乙方的开发进度进行确认和验收。

5.3. 原则上，甲方不得对本合同附件所约定的软件开发需求（要求）进行修改。甲方确需变更的，由甲乙双方另行协商开发进度、费用等事宜。

六、 乙方的权利与义务

6.1. 乙方保证对本合同项下软件拥有完整的知识产权，不存在侵犯他人知识产权的情形，否则由此产生的责任全部由乙方承担。

6.2. 乙方有权要求甲方按约付款，如甲方违反付款约定，乙方有权视情形中（终）止开发，并追究甲方相应的违约责任。

6.3. 除非双方另有书面约定或法律有强制性规定，乙方仅按照本合同的约定承担责任。

七、 违约责任

7.1. 一方违反本合同（包括附件）的约定给另一方造成损失的，应当赔偿另一方的损失，赔偿金额以本合同总金额为限，但本合同另有约定的除外。

八、 争议解决

8.1. 由本合同引起的或与本合同有关的任何争议，双方应当友好协商解决；如协商不成，双方均可向广州仲裁院申请仲裁，仲裁败诉方承担胜诉方发生的全部费用，包括但不限于仲裁费、保全费、律师费、交通费、鉴定费、差旅费等。

九、 其他

9.1. 合同一式四份，甲乙双方各执贰份，具有同等法律效力，经双方签章后生效。

9.2. 本合同所列附件为本合同不可或缺的部分，附件内容与本合同主文具有同等法律效力。

甲方签章：广州联展科技有限公司　　，乙方签章：华南农业大学

签章日期：2024.1.4　　签章日期：2024.1.4

附录

项目技术要求及参数

平台端	功能区域	功能
一、商城系统		
网站端 web 端/移动端（仅适配）	首页	展示精选的作品信息，并提供详情信息和购买入口
	列表页	展示所有作品信息，并提供详情信息入口
	详情页	展示对应作品的详细信息，及购买入口。
	购买页	对应作品的购买流程。
		支付：PayPal
	购物车页	加入购物车未进入购买流程的列表。
	注册/登录页	邮箱注册&登陆/账号字符注册&登陆
	用户信息页	展示已登录用户的信息，提供编辑功能。
	新闻列表页	展示新闻列表，可进入新闻详情。
	新闻详情页	H5 展示
	用户协议页	富文本
	联系我们页	关于我们&展示我们的介绍
后台管理系统	作品信息管理	创建作品、作品列表管理
	用户信息管理	用户列表（用户基本信息、注册时间、最后登陆时间、下单累计金额等）；用户详情（用户基本信息、订单记录）
	订单管理	全部订单、待付款订单、已付款订单、待发货订单、已发货订单、已完成订单、取消订单
	新闻管理	创建文章、筛选管理
	网站设置	logo 设置、banner 设置
		支付设置（PayPal）
	用户设置	用户协议管理
二、售后服务		
售后服务	一年合同质保期内的免费平台维护，团队远程微信、QQ、电话支撑;	

广东省大数据分析与处理重点实验室

开放基金项目结题报告

项目名称	招标采购信息抓取与格式化研究与应用
承担单位	华南农业大学
项目负责人	古万荣
起止时间	2017 年 01 月至 2018 年 12 月
联系电话	13719398640
电子邮箱	guwanrong@scau.edu.cn
填表时间	2018 年 12 月 24 日

项目信息	项目名称	招标采购信息抓取与格式化研究与应用		
	项目类别		研究/学科方向	**信息检索**
	获准金额	*1 万元*	起止时间	2017 年 01 月至 2018 年 12 月
项目负责人	姓 名	古万荣	职称	讲师
	所在单位	华南农业大学	身份证号/护照号	440229198201270018
	研究方向	信息检索与推荐		
预期成果和验收指标	论文一篇（三大索引） 专利一项			
研究内容总结	该项目是信息检索和推荐的应用型研究项目，信息抓取和格式化处理是互联网应用的一个重要方向，是互联网应用的数据基础和前提，可以为垂直搜索、个性化推荐等应用服务，具有广阔的应用前景和研究价值，本项目以此为基础，开发原型系统，并申请软件著作版权 2 项，专利 1 项。以此为基础，本项目研究还进行了扩展，对爬取和搜索算法应用到推荐和矩阵填补作用预测中，并发表相关论文 2 篇。			

成果列表

1、论文

序号	论文题目	发表刊物	作者	发表/出版时间	标注情况
1	基于Adaboost算法的药物--靶向蛋白作用预测算法	生物医学工程学杂志	古万荣，谢贤芬，何亦琛，张子烨	2018年12月	EI刊源，有标注
2	Optimization of Hypergraph based News Recommendation by Binary Decision Tree	International Conference on Intelligent Science and Big Data Engineering	Wanrong Gu, Xianfen Xie, Yijun Mao, Yichen He	2018.7	EI检索，有标注
3					
4					

2、专利

序号	专利名称	专利号/公开号	受理日期	专利类型	发明人
1	基于 DBN 的药物靶向蛋白作用预测算法	201810336860.0	2018.04.16	发明专利	古万荣，毛宜军，田绪红，黎嘉朗
2	基于标签路径和文本标点比特征融合的网页正文提取方法	201711061497 .8	2017.11.02	发明专利	黎嘉朗，古万荣，田绪红，毛宜军，李吉平
3					

成果列表

3、软件著作权

序号	软件名称	登记号	著作权人	发证日期
1	基于词向量模型的查询词扩展搜索平台	2018SR524456	华南农业大学	2018 年 7 月 6 日
2	基于 Nutch 的互联网信息抓取系统	2018SR525795	华南农业大学	2018 年 7 月 6 日
3				

4、获奖情况

序号	奖项名称	奖项等级	获奖时间	颁发单位
1				
2				
3				

人才培养情况	序号	研究生姓名	研究方向	论文题目	答辩时间
	1	黎嘉朗	生物信息	基于DBN的药物靶向蛋白作用预测研究	2018.06
	2	徐志文	信息检索	基于词向量的农业生产知识查询扩展研究	2018.06
	3				
其他					

项目负责人签字及审核意见：

项目负责人承诺：

我所承担的项目（编号：2017006，名称：招标采购信息抓取与格式化研究与应用）结题报告内容填写实事求是，数据详实。在今后的研究工作中，如有与本项目相关的成果，将标注“广东省大数据分析与处理重点实验室 广东 广州 510006（Guangdong Key Laboratory of Big Data Analysis and Processing,Guangzhou 510006, P.R.China）”；在基金资助栏注明受到本实验室开放基金资助,标注“广东省大数据分析与处理重点实验室开放基金项目资助”（Supported by the Opening Project of Guangdong Key Laboratory of Big Data Analysis and Processing）中英文字样+项目编号。

负责人（签章）：

日期：　　年　　月　　日

项目依托单位审查意见：

经办人（签章）：　　单位公章：

日期：　　年　　月　　日

重点实验室审核意见：

审核人（签章）：　　盖章：

日期：　　年　　月　　日

项目负责人签字及审核意见：

项目负责人承诺：

我所承担的项目（编号：2017006，名称：招标采购信息抓取与格式化研究与应用）结题报告内容填写实事求是，数据详实。在今后的研究工作中，如有与本项目相关的成果，将标注“广东省大数据分析与处理重点实验室 广东 广州 510006（Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou 510006, P.R.China）”；在基金资助栏注明受到本实验室开放基金资助，标注“广东省大数据分析与处理重点实验室开放基金项目资助”（Supported by the Opening Project of Guangdong Key Laboratory of Big Data Analysis and Processing）中英文字样+项目编号。

负责人（签章）：

日期：　　年　　月　　日

项目依托单位审查意见：

同意结题

经办人（签章）：

日期2018年12月29日

单位公章：

重点实验室审核意见：

审核人（签章）：　　盖章：

日期：　　年　　月　　日


科研空间
欢迎您，古万荣老师
科研动态
我的协公
我的项目
我的经费
我的成果
我的学术交流
我的考核
我的主页
常用下载
项目立项
项目申报（1）
所有项目
职称评定
主持的项目
参与的项目
博硕导评定
主持的项目
参与的项目
项目性质
纵向(15项)
横向(7项)
项目分类
教育部人文社科研究一般项目(3项)
全国统计科研计划项目(2项)
厅级其它项目(1项)
国家重点研发计划项目(1项)
其他纵向(1项)
实验室开放课题(1项)
纵向平台(1项)
科技计划项目(1项)
国家自然科学基金项目(1项)
2018年以前广东省自然科学基金(1项)
广东省社科基金一般项目(1项)
总数:10项 (表中经费单位：万元)
导出
打印参与
全选
序号
项目名称
评审等级
项目来源
主持人
合同经费（万元）
实到经费（万元）
结余经费（万元）
立项时间
结题时间
本人排名
项目课题组总人数
1 风险偏好、风险认知与农户小额信贷的"自我排斥"研究 A 教育部 陈思 8.0 8.0 8.0 2022-09-26 2025-12-01 3 5
2 交通无线磁阻传感器网络深度学习去噪方法研究 A 国家自然科学基金委员会 张足生 50.0 28.45 28.45 2024-12-01 2028-12-31 3 5
3 [illegible]贷目标偏移与政策优化研究 B 广东省哲学社会科学规划办公室 陈思 5.0 0.0 0.0 2021-10-14 2024-10-14 3 4
4 温氏股份基因数据处理算法研究 横向 毛宜军 15.0 15.0 15.0 2025-03-31 2025-05-30 3 4
5 知识驱动的科学数据智能分析软件系统与示范应用 A 科技部 黄琼 70.0 29.75 22.209874 2023-12-28 2027-12-28 4 10
6 深度学习训练数据生成服务 横向 王国华 8.0 8.0 2.88 2023-08-25 2023-11-25 4 4
7 融合知识库与语料库的汉语词义消歧研究 A 教育部 韦婷婷 8.0 0.0 0.0 2020-03-20 2022-12-01 5 5
8 一种针对表格字体的识别方法（专利权）合同 横向 王国华 4.0 4.0 4.0 2023-05-24 2023-11-24 5 5
9 渔业大数据系统标准化体系研究 其他纵向 黄琼 30.0 20.0 3.3333 2023-06-30 2024-06-30 5 10
10 基于移动式平台的水下场景立体重构与目标识别 B 厅局其它 黄琼 70.0 59.5 55.648433 2024-03-01 2026-02-28 7 10

检索证明

根据委托人提供的论文材料，委托人华南农业大学数学与信息学院古万荣（ Gu, Wanrong）1篇论文收录情况如下表。

序号	论文名称	发表刊物及发表的年月卷期/页码等	作者排名	论文等级	作者文中单位	收录情况	影响因子	中科院大类分区
1	An Improved Recommendation Based on Graph Convolutional Network	Journal of Intelligent Information Systems 出版年：2022 DEC 卷期：59 3 页码：801-823 文献类型：Article	通讯作者	B类	华南农业大学	SCI	IF2-year=2.504 IF5-year=2.507 (2021)	计算机科学 3区 Top期刊：否 (2022)

说明：论文等级和中科院大类分区按《华南农业大学学术论文评价方案（试行）》划分。

报告免责声明:如未盖章,报告无效

检索员：杨穗璇
华南农业大学图书馆
2023-03-13

华南农业大学图书馆 CAU LIBRARY

检索证明

根据委托人提供的论文材料，委托人华南农业大学数学与信息学院古万荣2篇论文收录情况如下表。

序号	论文名称	发表刊物及发表的年月卷期/页码等	作者排名	论文等级	作者文中单位	收录情况	影响因子	中科院大类分区
1	创新环境指数构建及基于自适应ARIMA的动态预警监控研究	统计与信息论坛 出版年：2021 卷期：36 6 页码：3-13 文献类型：	通讯作者	C类	华南农业大学	北大中文核心期刊2020		
2	基于多模型优化的超声图像肿瘤自动识别	计算机科学 出版年：2020 卷期：47 S1 页码：260 267 文献类型：	第一作者	B类	华南农业大学	北大中文核心期刊		

说明：论文等级和中科院大类分区按《华南农业大学学术论文评价方案（试行）》划分。

报告免责声明:如未盖章,报告无效

检索员：尹银怀
华南农业大学图书馆
2022-01-12

华南农业大学图书馆 CAULIB2022180012
华南农业大学图书馆 信息检索专用章

文 献 检 索 证 明

根据委托人提供的论文材料，委托人华南农业大学数学与信息学院，古万荣的2篇论文收录情况如下表。

序号	论文名称	发表刊物 及发表的年月卷期、页码等	作者排名	作者 文中单位	收录情况	文献类型
1	Optimization of Hypergraph Based News Recommendation by Binary Decision Tree	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), v 11266 LNCS, p 83-101, 2018	第一作者	华南农业大学	EI	Journal article (JA)
2	Drug-target protein interaction prediction based on AdaBoost algorithm	Shengwu Yixue Gongchengxue Zazhi/Journal of Biomedical Engineering, v 35, n 6, p 935-942, December 25, 2018 Language: Chinese	第一作者	华南农业大学	EI	Journal article (JA)

检索员：尹银怀

华南农业大学

信息检索专用章

2019年5月23日

检索证明

根据委托人提供的论文材料，委托人**华南农业大学数学与信息学院古万荣**10篇论文收录情况如下表。

序号	论文名称	发表刊物及发表的年月卷期/页码等	作者排名	论文等级	作者文中单位	收录情况	影响因子	中科院大类分区
1	Multi-factor Data Mining Analysis of Stock Index Volatility	ACM International Conference Proceeding Series 出版年：2021 May 28, 2021 卷期： 文献号：3470231 文献类型：Conference article (CA)	第四作者	/	华南农业大学	EI		
2	An Improved Muti-Task Learning Algorithm for Analyzing Cancer Survival Data	IEEE-ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS 出版年：2021 MAR-APR 卷期：18 2 页码：500-511 文献类型：Article	第一作者	B类	华南农业大学	SCI	IF2-year=3.71 IF5-year=3.395 (2020)	工程技术 3区 Top期刊：否 (2021)
3	Collaborative Score-Prediction Method for Non-Random Missing Data	Huanan Ligong Daxue Xuebao /Journal of South China University of Technology (Natural Science) 出版年：2021 January 2021 卷期：49 1 页码：47-57 文献类型：Journal article (JA)	第一作者	B类	华南农业大学	EI		
	High Accuracy Drug-Target	Proceedings of 2020 IEEE International Conference on Power, Intelligent Computing						

华南农业大学图书馆 SCAU LIB2022101163

华南农业大学图书馆 信息检索专用章

4	Protein Interaction Prediction Method based on DBN	and Systems, ICPICS 2020 出版年：2020 July 2020 卷期： 文献号：9202120 文献类型：Conference article (CA)	第一作者	/	华南农业大学	EI		
5	High Accuracy Drug-Target Protein Interaction Prediction Method based on DBN	Proceedings of 2020 IEEE International Conference on Power, Intelligent Computing and Systems, ICPICS 2020 出版年：2020 July 2020 卷期： 文献号：9202120 文献类型：Conference article (CA)	第一作者	/	华南农业大学	EI		
6	Improved Transfer Learning Algorithm Based on Cross-domain in Recommendation System	Huanan Ligong Daxue Xuebao / Journal of South China University of Technology (Natural Science) 出版年：2020 November 2020 卷期：48 11 页码：99-106 文献类型：Journal article (JA)	通讯作者	B类	华南农业大学	EI		
7	High Accuracy Thyroid Tumor Image Recognition Based on Hybrid Multiple Models Optimization	IEEE ACCESS 出版年：2020 卷期：8 页码：128426-128439 文献类型：Article	第一作者	B类	华南农业大学	SCI	IF2-year=3.367 IF5-year=3.671 (2020)	计算机科学 3区 Top期刊：否 (2020)
		2020 IEEE 3rd International Conference on Automation, Electronics and Electrical						

华南农业大学图书馆 LIB20220210163
华南农业大学图书馆 信息检索专用章

8	Drug-protein Interaction Prediction based on Improved AdaBoost Algorithm	Engineering, AUTEEE 2020 出版年：2020 November 20, 2020 卷期： 文献号：9315654 文献类型：Conference article (CA)	第一作者	/	华南农业大学	EI		
9	A collaborative-filtering recommendation based on the trust network	Proceedings - 2019 International Conference on Intelligent Computing, Automation and Systems, ICICAS 2019 出版年：2019 December 2019 卷期： 文献号：9051291 文献类型：Conference article (CA)	第二作者	/	华南农业大学	EI		
10	Optimization of hypergraph based news recommendation by binary decision tree	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 出版年：2018 2018 卷期：11266 LNCS 页码：83-101 文献类型：Conference article (CA)	第一作者	/	华南农业大学	EI		

说明：论文等级和中科院大类分区按《华南农业大学学术论文评价方案（试行）》划分。

华南农业大学图书馆 SCAU LIB220221016

华南农业大学图书馆 信息检索专用章

报告免责声明:如未盖章,报告无效

检索员：刘汉忠
华南农业大学图书馆

IEEE/ACM Transactions on Computational Biology and Bioinformatics

3.4 Impact Factor	0.00856 Eigenfactor
0.799 Article Influence Score	9.7 CiteScore Powered by Scopus

Title History

Aims & Scope

IEEE/ACM Transactions on Computational Biology and Bioinformatics emphasizes the algorithmic, mathematical, statistical, and computational methods that are central in bioinformatics and computational biology; the development and testing of effective computer programs in bioinformatics; the development of biological databases; and important biological results that are obtained from the use of these methods, programs and databases; the emerging field of Systems Biology, where many forms of data are used to create a computer-based model of a complex biological system. The publication represents a mixture of three research modalities: a) fundamental methodological, algorithmic, mathematical and statistical research directly motivated by biological issues; b) papers focusing on experimental and implementation issues; and c) papers on serious application of methods and programs that lead to discoveries of biological significance. Increasingly, papers contain elements of all three modalities. Specific topics of interest include, but are not limited to, sequence analysis, comparison and alignment methods; motif, gene and signal recognition; molecular evolution; phylogenetics and phylogenomics; determination or prediction of the structure of RNA and Protein in two and three dimensions; DNA twisting and folding; gene expression and gene regulatory networks; deduction of metabolic pathways; micro-array design and analysis; proteomics; functional genomics; molecular docking and drug design; computational problems in genetics such as linkage and QTL analysis, linkage disequilibrium analysis in populations, and haplotype determination; systems biology.

This title ceased publication in 2024. The current retitled publication is IEEE Transactions on Computational Biology and Bioinformatics.

Publication Fees

This publication is hybrid, meaning that it is supported by subscriptions and applicable article processing charges (APCs). Although there is no cost for publishing with IEEE, authors may wish to take advantage of some of our fee-based offerings, for example, choosing to publish open access in a hybrid journal. Visit the IEEE Author Center for more information on available options. The open access APC for this title is **$2,495 USD**. Learn more

Popular | Recent | Most Cited

Alignment-Free Sequence Comparison: A Systematic Survey From A Machine Learning Perspective
Katrin Sophie Bohnsack; Marika Kaden; Julia Abel; Thomas Villmann

Modular Multi–Source Prediction Of Drug Side–Effects With DruGNN
Pietro Bongini; Franco Scarselli; Monica Bianchini; Giovanna Maria Dimitri; Niccoló Pancino; Pietro Lió

Explainable Drug Repurposing Approach From Biased Random Walks
Filippo Castiglione; Christine Nardini; Elia Onofri; Marco Pedicini; Paolo Tieri

A Fitted Sparse-Group Lasso For Genome-Based Evaluations
Jan Klosa; Noah Simon; Volkmar Liebscher; Dörte Wittenburg

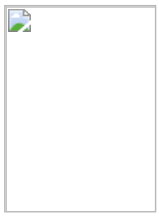
EnsDeepDP: An Ensemble Deep Learning Approach For Disease Prediction Through Metagenomics
g Shen; Jinlin Zhu; Zhaohong Deng; Wenwei Lu; Hongchao Wang

Show All

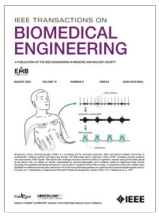
Related Magazines or Journals



IEEE Reviews in Biomedical Engineering



IEEE Transactions on Information Technology in Biomedicine



IEEE Transactions on Biomedical Engineering

Submission Guidelines

Become a Reviewer

Open Access Publishing Options

Add Title To My Alerts Add to My Favorites

Contact Information
Editor-in-Chief

*

Author Center

Publishing Policies

© Copyright 2025 IEEE - All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies.

IEEE/ACM Transactions on Computational Biology and Bioinformatics

3.4
Impact Factor

0.00856
Eigenfactor

0.799
Article Influence Score

9.7
CiteScore
Powered by Scopus

Title History

Issue 2 • March-April-2021

Get Entire Issue Now
Purchase or Sign in

Back to navigation

Search within results

Items Per Page | Export | Email Selected Results

Showing **1-25** of **39**

Author

Affiliation

Supplemental Items

A Cascade of 2.5D CNN and Bidirectional CLSTM Network for Mitotic Cell Detection in 4D Microscopy Image
Titinunt Kitrungrotsakul; Xian-Hau Han; Yutaro Iwamoto; Satoko Takemoto; Hideo Yokota; Sari Ipponjima; Tomomi Nemoto; Wei Xiong; Yen-Wei Chen
Publication Year: 2021 , Page(s): 396 - 404
Cited by: Papers (11)
Abstract HTML

A Fast Linear Neighborhood Similarity-Based Network Link Inference Method to Predict MicroRNA-Disease Associations
Wen Zhang; Zhishuai Li; Wenzheng Guo; Weitai Yang; Feng Huang
Publication Year: 2021 , Page(s): 405 - 415
Cited by: Papers (83)
Abstract HTML

A Faster Algorithm for Computing the Kernel of Maximum Agreement Subtrees
Biing-Feng Wang; Krister M. Swenson
Publication Year: 2021 , Page(s): 416 - 430
Cited by: Papers (2)
Abstract HTML

A Gene Rank Based Approach for Single Cell Similarity Assessment and Clustering
Yunpei Xu; Hong-Dong Li; Yi Pan; Feng Luo; Fang-Xiang Wu; Jianxin Wang
Publication Year: 2021 , Page(s): 431 - 442
Cited by: Papers (10)
Abstract HTML

A Hybrid Supervised Approach to Human Population Identification Using Genomics Data
Sahar Araghi; Thanh Nguyen
Publication Year: 2021 , Page(s): 443 - 454
Cited by: Papers (5)
Abstract HTML

A Novel Branch-and-Bound Algorithm for the Protein Folding Problem in the 3D HP Model

Hsin-Hung Chou; Ching-Tien Hsu; Li-Hsuan Chen; Yue-Cheng Lin; Sun-Yuan Hsieh
Publication Year: 2021 , Page(s): 455 - 462
Cited by: Papers (1)

Abstract HTML

A Novel Drug Repositioning Approach Based on Collaborative Metric Learning
Huimin Luo; Jianxin Wang; Cheng Yan; Min Li; Fang-Xiang Wu; Yi Pan
Publication Year: 2021 , Page(s): 463 - 471
Cited by: Papers (12)

Abstract HTML

An Analytical Review of Computational Drug Repurposing
Seyedeh Shaghayegh Sadeghi; Mohammad Reza Keyvanpour
Publication Year: 2021 , Page(s): 472 - 488
Cited by: Papers (31)

Abstract HTML

An Evolutionary Profile Guided Greedy Parallel Replica-Exchange Monte Carlo Search Algorithm for Rapid Convergence in Protein Design
Anupam Banerjee; Kuntal Pal; Pralay Mitra
Publication Year: 2021 , Page(s): 489 - 499
Cited by: Papers (7)

Abstract HTML

An Improved Muti-Task Learning Algorithm for Analyzing Cancer Survival Data
Wanrong Gu; Ziye Zhang; Xianfen Xie; Yichen He
Publication Year: 2021 , Page(s): 500 - 511
Cited by: Papers (6)

Abstract HTML

ANCA: Alignment-Based Network Construction Algorithm
Kevin Chow; Aisharjya Sarkar; Rasha Elhesha; Pietro Cinaglia; Ahmet Ay; Tamer Kahveci
Publication Year: 2021 , Page(s): 512 - 524
Cited by: Papers (17)

Abstract HTML

Bayesian Inference in Y-Linked Two-Sex Branching Processes with Mutations: ABC Approach
Miguel González; Cristina Gutiérrez; Rodrigo Martínez
Publication Year: 2021 , Page(s): 525 - 538
Cited by: Papers (3)

Abstract HTML

CNV_IFTV: An Isolation Forest and Total Variation-Based Detection of CNVs from Short-Read Sequencing Data
Xiguo Yuan; Jiaao Yu; Jianing Xi; Liying Yang; Junliang Shang; Zhe Li; Junbo Duan
Publication Year: 2021 , Page(s): 539 - 549
Cited by: Papers (36)

Abstract HTML

Colored Network Motif Analysis by Dynamic Programming Approach: An Application in Host Pathogen Interaction Network
Sourav Biswas; Sumanta Ray; Sanghamitra Bandyopadhyay
Publication Year: 2021 , Page(s): 550 - 561
Cited by: Papers (3)

Abstract HTML

Deep Convolutional and Recurrent Neural Networks for Cell Motility Discrimination and Prediction

Jacob C. Kimmel; Andrew S. Brack; Wallace F. Marshall
Publication Year: 2021 , Page(s): 562 - 574
Cited by: Papers (28)

Abstract HTML

DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines
Min Li; Yake Wang; Ruiqing Zheng; Xinghua Shi; Yaohang Li; Fang-Xiang Wu; Jianxin Wang
Publication Year: 2021 , Page(s): 575 - 582
Cited by: Papers (84)

Abstract HTML

Improve the Resolution and Parallel Performance of the Three-Dimensional Refine Algorithm in RELION Using CUDA and MPI
Jingrong Zhang; Zihao Wang; Zhiyong Liu; Fa Zhang
Publication Year: 2021 , Page(s): 583 - 595
Cited by: Papers (6)

Abstract HTML

iPhosH-PseAAC: Identify Phosphohistidine Sites in Proteins by Blending Statistical Moments and Position Relative Features According to the Chou's 5-Step Rule and General Pseudo Amino Acid Composition
Muhammad Awais; Waqar Hussain; Yaser Daanial Khan; Nouman Rasool; Sher Afzal Khan; Kuo-Chen Chou
Publication Year: 2021 , Page(s): 596 - 610
Cited by: Papers (51)

Abstract HTML

MCHMDA:Predicting Microbe-Disease Associations Based on Similarities and Low-Rank Matrix Completion
Cheng Yan; Guihua Duan; Fang-Xiang Wu; Yi Pan; Jianxin Wang
Publication Year: 2021 , Page(s): 611 - 620
Cited by: Papers (28)

Abstract HTML

MGRFE: Multilayer Recursive Feature Elimination Based on an Embedded Genetic Algorithm for Cancer Classification
Cheng Peng; Xinyu Wu; Wen Yuan; Xinran Zhang; Yu Zhang; Ying Li
Publication Year: 2021 , Page(s): 621 - 632
Cited by: Papers (54)

Abstract HTML

Object Weighting: A New Clustering Approach to Deal with Outliers and Cluster Overlap in Computational Biology
Alexandre Gondeau; Zahia Aouabed; Mohamed Hijri; Pedro R. Peres-Neto; Vladimir Makarenkov
Publication Year: 2021 , Page(s): 633 - 643
Cited by: Papers (16)

Abstract HTML

Optimal Bayesian Transfer Learning for Count Data
Alireza Karbalayghareh; Xiaoning Qian; Edward R. Dougherty
Publication Year: 2021 , Page(s): 644 - 655
Cited by: Papers (1)

Abstract HTML

Physically-Inspired Gaussian Process Models for Post-Transcriptional Regulation in Drosophila
Andrés F. López-Lopera; Nicolas Durrande; Mauricio A. Álvarez
Publication Year: 2021 , Page(s): 656 - 666
Cited by: Papers (5)

Abstract HTML

Predicting in-vitro Transcription Factor Binding Sites Using DNA Sequence + Shape
Qinhu Zhang; Zhen Shen; De-Shuang Huang
Publication Year: 2021 , Page(s): 667 - 676
Cited by: Papers (50)

Abstract HTML

Predicting the Antigenic Relationship of Foot-and-Mouth Disease Virus for Vaccine Selection Through a Computational Model
Jingxuan Qiu; Tianyi Qiu; Qingli Dong; Dongpo Xu; Xiang Wang; Qi Zhang; Jing Pan; Qing Liu
Publication Year: 2021 , Page(s): 677 - 685
Cited by: Papers (7)

Abstract HTML

Submission Guidelines

Become a Reviewer

Open Access Publishing Options

Add Title To My Alerts Add to My Favorites

Contact Information
Editor-in-Chief
*

Author Center

Publishing Policies

IEEE Personal Account
CHANGE USERNAME/PASSWORD

Purchase Details
PAYMENT OPTIONS
VIEW PURCHASED DOCUMENTS

Profile Information
COMMUNICATIONS PREFERENCES
PROFESSION AND EDUCATION
TECHNICAL INTERESTS

Need Help?
US & CANADA: +1 800 678 4333
WORLDWIDE: +1 732 981 0060
CONTACT & SUPPORT

Follow

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | IEEE Ethics Reporting | Sitemap | IEEE Privacy Policy
A public charity, IEEE is the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity.

© Copyright 2025 IEEE - All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies.

An Improved Muti-Task Learning Algorithm for Analyzing Cancer Survival Data

Wanrong Gu, Ziye Zhang, Xianfen Xie, and Yichen He

Abstract—Survival analysis is a popular branch of statistics. At present, many algorithms (like traditional multi-tasking learning model) cannot be applied well in practice because of censored data. Although using some model (like parametric regression model) can avoid it, they need strict assumptions. This undermines the very nature of things, which is very detrimental to the study of practical problems. The method proposed in this paper can apply well to the censored data, but does not need to make any additional assumptions about the original problem. It can be said that it breaks through the above two kinds of major limitations. The algorithm is a kind of inductive transfer learning method, which can fully obtain the information in the censored data, using domain-specific information implicit in each feature to enhance the generalization capability of the model. We also used two common performance metrics as criteria to judge the predictive performance differences between the models in this article and those of other mainstream models. The results show that the model proposed in this paper is 10 $\sim$ 15 percent higher than other mainstream models, which proves that our multi-task learning model has a great advantage in the survival analysis of cancer genes.

Index Terms—Cancer genes, multi tasking, ADMM, survival analysis

1 Introduction

Today's social and economic development has raised the level of material consumption, and all kinds of cancer caused by environmental pollution has become the most important disease that endangers human life. Factors such as living environment and living habits all have great influence on the incidence of cancer. With the continuous progress of medical technology, the survival time of cancer patients has been improved. But the survival of most cancer patients remains grim.

Survival analysis is a branch of the mathematical statistics that emerged in the 1970s. It can analyze the relationship between survival time and influencing factors by making a simple prediction of things in various fields. Simple survival analysis are usually related binary classification events, which can predict a patient's survival at a specific point in time. And further survival analysis studies can help doctors guide the treatment and evaluate whether the treatment is appropriate for patients.

Survival analysis was originally applied in bio-medicine and bio-information engineering. It can analyze the relationship between survival time and influencing factors by simply predicting things in various fields. Therefore, survival analysis in modern scientific research is not only used in the field of medicine, but also widely used in finance, insurance actuarial science, engineering, economics and other disciplines. With the development of the research work, survival analysis should also be constantly improved. It is necessary to record the survival time of the study object and to observe the change of state influenced by various factors in the whole survival process. Usually, the survival time of objects does not conform to a normal distribution but some other distribution like exponential distribution. And most of the data contained the censored data. So it requires researchers to build a model of survival analysis that maximizes the use of information in the data and uses training models to pick out the main factors that affect survival time. In the study, the error of the model caused by the censored data needs to be reduced to the extent possible [9], [22], in order to improve the generalization of survival analysis models.

This paper proposes an Improved Multi-task Learning Algorithm for Analyzing Cancer Survival Data (IMACS), which defines the original problem of survival time as multi-task learning problem. The primary purpose of using multitasking is that the multitasking algorithm can learn the shared representation of the task and reduce the predictive error of each task. At the same time, this paper divides the real time of existence in the living data into a smaller integral time interval, providing a more accurate representation of the state of the matter at the beginning of each interval. Therefore, the method of using this article can predict the more accurate survival time of each instance.

In this paper, another advantage of multi-task learning for survival time prediction is that it can transform the regression into a series of related dichotomous problems. The corresponding classifier only needs to focus on modeling local problems at each time interval, and thus have a more accurate predictive ability than a regression model that only simulates once. In addition to the linear hypothesis, the model in this

- *W. Gu and Y. He are with the School of Mathematics and Information, South China Agricultural University, Guangzhou 510642, P.R. China. E-mail: guwanrong@scau.edu.cn, heyichen666@hotmail.com.*
- *Z. Zhang is with the School of Mathematical, South China University of Technology, Guangzhou 510660, China. E-mail: ma_zere_x1@mail.scut.edu.cn.*
- *X. Xie is with the School of Economy, Jinan University, Guangzhou 510632, P.R. China. E-mail: txiexianfen2009@jnu.edu.cn.*

Manuscript received 27 Jan. 2019; revised 26 May 2019; accepted 2 June 2019. Date of publication 5 June 2019; date of current version 1 Apr. 2021. (Corresponding author: Ziye Zhang.) Digital Object Identifier no. 10.1109/TCBB.2019.2920770

1545-5963 © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: South China Agricultural Univ. Downloaded on June 18,2021 at 08:21:57 UTC from IEEE Xplore. Restrictions apply.

paper only assumes the multi-task learning model, and does not propose any additional assumptions other than the model, that is to say, the data feature is linear with the target function. This ensures the accuracy of the model.

In the research of survival analysis, censored data is inevitable because of the concentration of cancer genetic data. But in response to these censored data, this paper uses a novel approach – Generating Analysis models that build an additional matrix of instructions for all instances, the analysis found that the characteristics of the survival data in the state of existence (once the state of the case changes, it does not revert to its original state, in which the patient can not change from a state of death to a living state).

By handling the survival data skillfully, the model is able to learn simultaneously between censored and normal data. Besides, the paper introduce Frobenius norm and the $L_{2,1}$ norm as additional regularization items for multi-task learning models. Using the Frobenius norm as an additional regularization term can alleviate the problem of overfitting of high-dimensional data. And using the $L_{2,1}$ norm as another additional regularization term makes multiple predicted variables to share similar sparse patterns, then select important hidden features and reduce the forecast errors for each task. Other than that, using the $L_{2,1}$ norm can also select some common features between instances to help the model process the galway dataset effectively.

In the algorithm optimization phase of the model, an alternate direction multiplier algorithm [3] is used to solve the convex optimization problem. It combines the excellent convergence properties of the Dual ascent with the excellent convergence in the Method of multiplier. It can decompose the original function and the amplification function at the same time, introduce new variables and optimize alternately by crossing direction. Furthermore, it transforms the constrained optimization problem of data features into a Euclid projection problem that can be effectively learned.

Finally, this paper introduces the source and data content of several high-dimensional cancer gene expression survival data sets, and uses different algorithms, including the algorithms presented in this paper and several main algorithms, to analyze the data. Then we compare the predictive performance of the algorithms through two kinds of evaluation indexes which are C-index and weighted average value of AUC. Finally, the paper also testes scalability and stability of the model by generating random data sets.

2 Related Work

In recent years, more and more researchers have used the methods of survival analysis and machine learning to carry out various kinds of medical research [12], [13]. In the field of survival analysis of cancer genes, a large number of researchers have conducted targeted experiments on different research goals and summarized many methods [5], [10], [15], [17], [18], [23], their methods will be mentioned below.

2.1 The Research Methods of Survival Analysis

The common research methods of survivorship analysis can be divided into three kinds: parametric method, non-parametric method and semi-parametric method.

2.1.1 Parametric Method

Usually, this method is to analyze the data of the sample and to assume that the lifetime is subject to a specific distribution based on sample data. At the same time, the parameters required are estimated in the constructed model. Based on the characteristics of known distribution, the survival time of the observed samples is analyzed to obtain the probability distribution model of survival time. The research method needs to assume an effective model in the input space and summarize the survival analysis into estimating the survival function by estimating the parameters in the sample. The parametric method needs to analyze the survival data in advance, and deduce the essential law reflected in the data. The hypothetical distribution includes exponential distribution, log-normal distribution, Weibull distribution, Logistic distribution, Gamma distribution, etc. The parametric method can analyze the influence of risk factors on survival time in the survival function by comparing the function of two or more groups of survival time distribution functions, thus establishing the relationship between survival time and risk factors. This is also the advantage of the parametric method, and the disadvantage is the need to obtain the distribution of survival time in advance.

2.1.2 Semi-Parametric Method

It combines the advantages of the parametric method with the non-parametric method [17]. It does not need to know the distribution of lifetime information and the base risk rate to estimate the regression coefficient and the risk function. The semi-parametric method needs to establish a regression model which combines the parametric model and the non-parametric model to analyze the distribution law of survival time. At the same time, the risk factor is used as a covariant to analyze its effect on survival time. Therefore, the semi-parametric method is the most widely used method in the field of survival analysis. And the most common semi-parametric methods are Cox proportional risk regression model [5] and linear regression model. Cox Proportional risk regression model is one of the earliest and most widely used methods of survival analysis. The model can analyze the influence of many risk factors on the survival time, and also can deal with the censored data. His method has a great influence on this paper.

In Cox proportional risk regression model, survival rate is expressed as $P(T \geq t)$, indicating the ratio of those who survived at the start to those who survived at the end of the study. And it can also be expressed as $P = 1-\mathrm{q}$, where q is mortality Rate. The survival function represents the relationship of the survival rate with time t. It is noted as $S(t)$. The survival rate at time t is the functional value of the survival function at time t.

The risk function is the probability that the subject will die within the period $t \sim t + \Delta t$ if he is still alive at time t, which is expressed as $h(t)$. It also shows the probability of death of the subject who is still alive at time t. Let $X = (X_1, X_2, \ldots, X_p)$ to be a covariates, which is usually a feature of the data. So we have

$$h(t, X) = \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t | T \geq t)}{\Delta t}. \quad (1)$$

Authorized licensed use limited to: South China Agricultural Univ. Downloaded on June 18,2021 at 08:21:57 UTC from IEEE Xplore. Restrictions apply.

In the formula, T is the survival time of the observed. So the relationship between risk function and survival function can be expressed as follow:

$$S(t, X) = \exp\left[-\int_0^t h(u, X)du\right]. \quad (2)$$

The main aim of survival analysis is to study the relationship between covariates X and survival function $S(t, X)$. When $S(t, X)$ is affected by many risk factors, that is, $X = (X_1, X_2, \ldots, X_p)$ is the vector, the traditional method is to consider using regression equation to solve the problem. But the time variable t can not satisfy the requirement of normal distribution and homogeneous variance because of the large amount of censored data in survival analysis. So it can not be solved by traditional regression method.

However, the Cox model does not directly consider the relationship between covariates X and $S(t, X)$, but uses $h(t, X)$ as dependent variable to solve the problem. Suppose that $h(t, X)$ is the risk function associated with the covariates X at time t, and $h_0(t)$ is the base risk function. $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is the partial regression coefficient, which is estimated from the actual data. Then we have

$$h(t, X) = h_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p) \quad (3)$$

$$\ln\frac{h(t, X)}{h_0(t)} = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p. \quad (4)$$

In the formula, $\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ is called prognostic index. Usually, the Cox model can not get the value of $h_0(t)$ directly. In many cases, only the parameter β needs to be estimated. Even if $h_0(t)$ is unknown, the parameter β can still be estimated by parameter analysis and other methods. Because the model has the parameter β and the risk function $h_0(t)$ has not been defined, the risk function $h_0(t)$ needs to be assumed. Therefore, this model is not a complete parametric model, but a semi-parametric model.

In the study of survival analysis, the biggest problem with using the Cox proportional risk model is how to deal with the censored data and high-dimensional data. A good solution would be to introduce a baseline risk function or a maximum likelihood value that uses the likelihood function [10], [23], which can handle data sets with censored data. When processing some high-dimensional data, the Cox model can also be combined with some regularization methods to enhance its effectiveness.

2.1.3 Non-Parametric Method

The method needs to assume that the type of the object studied is known, but the form of the overall probability density function is unknown. It does not require the distribution of survival time, but only needs to evaluate the survival rate according to the order of the sample provided. And it uses non-parametric test method [15] to analyze the influence of risk factors on survival time. Non-parametric methods include the life table method, the product limit estimate (Kaplan-meier Method), and so on. This method is more suitable for the case where the parameters have more overall distribution information, so it is more suitable for the time when there are more data samples.

Kaplan-Meier estimation [18] is the earliest and most widely used method of time analysis. In survival analysis in medicine, there are many factors that affect a patient's death, the most common such as age, gender, height, weight, blood pressure, lifestyle, duration of illness, etc. The effect of different factors on the disease is not consistent, and according to the influence factors, the survival analysis is divided into single factor and multi-factor analysis. The common single-factor survival analysis algorithm has Kaplan-Meier method and survival table method. This kind of algorithm can not depend on other covariate, and can analyze the influence of a certain influence factor on the survival time without changing the other covariate. But the single-factor survival analysis algorithm can only judge the effect of the influence factor on the survival time, and can not know whether the effect is really influenced by this factor, how much impact, multi-factor survival analysis is needed to solve the problem at this time. Multivariate survival analysis algorithms such as the Cox proportional risk regression model can analyze the influence of many factors on survival time, and can find out the effect of all the factors on the survival time, select the most important influencing factors for targeted treatment. The combined Kaplan-Meier method and Cox proportional risk regression model can get better results than Kaplan-Meier or Cox model alone. Kaplan-meier method can be used to analyse therapeutic outcomes while achieving survival and mortality, but its limited ability to analyse the impact of survival is reflected in its inability to make direct life-time predictions.

The advantage of the non-parametric method is that it does not require the distribution of survival time, but the disadvantage is that it can not establish the relationship between the survival time and the risk factors.

2.2 Other Regression Models

2.2.1 Parametric Review Regression Model

It appears as an alternative to the Cox model. International research into the review of regression models has been progressing since the Tobin [28]. The regression model includes parametric review regression model, non-parametric review regression model and semiparametric review regression model, the most standard setting form is parameter review regression model. The model assumes that the Logarithm of the lifetime or lifetime of all instances in the data follows a specific distribution, while the random error term follows the standard assumption that the mean value is zero, the variance constant, and the independent normal distribution. This set-up is the most basic condition to be able to review the regression model using parameters. There are a number of ways to review the regression model with this set of parameters, such as maximum likelihood, weighted least squares, etc. Jl Horowitz and GR Neumann [16] systematically analyzed the parameters model and semi-parametric model, and compared and analyzed the model selection, data type, model test and so on. At the same time, the parameter review regression model added penalty items [20] can also handle high-dimensional censored data.

2.2.2 Linear Regression Models

Linear regression model is also an important branch of survival analysis. Strictly speaking, linear regression is a special

Authorized licensed use limited to: South China Agricultural Univ. Downloaded on June 18,2021 at 08:21:57 UTC from IEEE Xplore. Restrictions apply.

parameter to review the regression method, and the reason of separating linear regression model in data analysis is because linear regression is a basic and important method [26]. Because of the censored data in the analysis of survival data, least squares can not be directly applied in survival analysis studies [24]. Tobin's Tobit model [28] was one of the first attempts to use linear regression to process the censored data that appeared in the survival analysis. Later, El Kaplan proposed using the Buckley-James (BJ) model to solve the problem of survival analysis by combining the Kaplan-meier model in the non-parametric model. Recently, the Buckley-James model (EN-BJ), which uses Elastic Net technology of Elastic Net [30], combined with the weighted linear model, can also effectively deal with the problem of survival analysis with high-dimensional data [1].

2.3 Multi-Task Learning Algorithm

As soon as the multi-task learning model is put forward, it has obtained a great response in various university fields. Multi-task learning (MTL) is a kind of migration learning method in machine learning. Caruana R et Al. introduced the concept of MTL for the first time [4], another task is improved by using the information contained in the training signals contained in other related tasks. In this paper, he introduces the learning methods of multi-task learning in detail, and in dozens of cases, experiments have shown that multitasking adds shared representation to neural network tasks to get better models, such findings have inspired more follow-up research. The T Evgeniou and others have created a formalized framework for multi-task learning [6], and the regularization programme can easily integrate tasks into existing learning formulas for coupling multi-task learning, thus providing flexible extensions for existing algorithms, multi-task learning is the solution to this problem. From the point of view of machine learning, multi-task learning can be regarded as a kind of induction migration. Induction migration improves the model by introducing a inductive bias, making the model more prone to certain assumptions. The common reductive bias is L_1 regularization, which makes the multi-task learning model more likely to produce sparse solutions.

Multi-tasking learning mode puts the censored data into the multi-task learning model together with the uncensored data, so that the model can obtain the information of both data. At the same time, sharing representations between related features makes the multi-task learning model better summarize the original cancer gene survival analysis tasks. So this paper chooses to use multi-task learning model to study survival analysis.

In recent years, multi-task learning model has been applied more and more widely, and combined with other algorithms of machine learning, the experimental results are more excellent. In the traditional medical field, the study of multi-task learning is more mature. In bio-graphic analysis, Zhang and Zeng combined multiple task learning strategies [7], [11], [14], [32], the depth convolutional neural network trained on large natural image sets serve as feature extractors for the ISH image and then fine-tuned by a multi-task classification of the training model with the ISH icon, the middle layer of the deep model produces the best representation of gene expression patterns. At the same time, multi-task learning is also widely used in the areas of HIV prediction and treatment [2], [25] and cancer genes research [29], and in the direction of the prognosis [33]. In other areas, based on a multi-task framework, Stephane Fotso proposes the Neural Multi-Task Logistic Regression (N-MTLR) model [8], which makes performance of the algorithm further improve.

3 Optimization Algorithm Based on Multi-Task Learning Model

By analyzing the relevant information in the survival data, the censored data is processed and put into the multi-task learning model together with the normal data, so that the model can obtain the information of both data simultaneously. At the same time, we can share the representation of related features so that the multi-task learning model used in this paper can better summarize the original task of cancer gene survival analysis.

3.1 Matrix Norm

In the research of survival analysis, there will inevitably be a large amount of high-dimensional data. Galway data usually refers to data with a high number of characteristics and a small number of patients. Because of the appearance of high-dimensional data, the training time of the model increases linearly. Only when the data dimension is reduced to less than the sample quantity can the model be more stable and reliable.

The usual way to reduce the data dimension is to introduce a norm penalty. Because the model used in this paper is a multi-task learning model, the Frobenius norm penalty term and the $L_{2,1}$ norm penalty term can be used to reduce the fitting degree of the model and reduce the data feature dimension. Both the Frobenius norm and the $L_{2,1}$ norm are matrix norm, which is often used to solve the optimization problem of the matrix.

3.1.1 $L_{2,1}$ Norm Penalties

In fact, the $L_{2,1}$ norm is a matrix norm. The $L_{2,1}$ norm of the matrix can be calculated by calculating the sum of the L_2 norm for each row in the target matrix. When solving the minimization problem, only the sum of the L_2 norm of all rows of the matrix is the minimum that minimizes the objective function. The minimum value of the L_2 norm in each row means that the norm can be minimized when as many values as possible appear in the row, and each row takes as many as 0 values as is a sparse definition. Assuming that $(a_1, \ldots, a_n)$ is the n-column vector of matrix A, the $L_{2,1}$ norm is the sum of the L_2 (Euclid norm) of all columns in the matrix, expressed as

$$\|A\|_{2,1} = \sum_{j=1}^{n} \|a_j\|_2 = \sum_{j=1}^{n} \left(\sum_{i=1}^{m} |a_{ij}|^2 \right)^{\frac{1}{2}}. \tag{5}$$

3.1.2 Frobeniusnorm Penalties

The Frobenius norm is actually an $L_{2,2}$ norm, which is also a matrix norm. The Frobenius norm penalty is often used to ease the overfitting of high-dimensional data, usually in the form of

Authorized licensed use limited to: South China Agricultural Univ. Downloaded on June 18,2021 at 08:21:57 UTC from IEEE Xplore. Restrictions apply.

$$\|A\|_F = \|A\|_{2,2} = \left(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2\right)^{\frac{1}{2}} = \left(\sum_{i=1}^{\min\{m,n\}}\sigma_i^2(A)\right)^{\frac{1}{2}}. \tag{6}$$

The implication of over fitting is that the model overtrains the training data so that the model works well in the training data, and the performance on the test data is not good. However, the actual occurrence of fitting is caused by the combination of two factors: data and models. The training data may not be representative of the overall situation in the sample, and may even differ significantly from the overall situation. The complexity of the training models can not represent the overall data, resulting in greater training errors, to create an overfitting phenomenon.

The purpose of the L_2 norm is to increase the differences among the features, while the Frobenius norm not only increases the difference of features within the column, but also increases the characteristic differences within the row. By adding the Frobenius norm to the multi-task learning formula, we can get a better result of reducing the fit.

3.2 Alternating Direction Method of Multipliers

Alternating Direction Method of Multipliers (ADMM) is a kind of algorithm which combines the decomposition of the dual-ascending method with the upper bound convergence property of the multiplier method, which is often used in the optimization of convex functions. In a paper by Stephen Boyd [3], the author has systematically analyzed the ADMM Algorithm, gradually extrapolating the formula, which is available in the author's article.

4 Survival Analysis Based on Multi-Task Learning Model

This section optimizes the original multi-task learning model with non-negative non-increment and time smoothing. Unlike traditional multi-task learning models, the processed censored data can still be used for training. In addition, two matrix penalty items were added to solve the problem of sparse feature and reduction model. In the optimization of the model Algorithm, the alternating-direction multiplier (ADMM) is introduced to transform the convex optimization problem of the target function into the optimization of two iterative subfunctions, and the improved algorithm of this paper is proposed. Then we analyze the convergence and complexity of the model in detail.

4.1 Modeling and Optimization

In the actual problem of survival analysis, the processing of the survival data is a very important step. For each patient, not only do you need to obtain data on a patient's characteristics, you also need to get the corresponding time to live. Because the patient's survival data is missing data, the time it takes to survive at maximum observation time is not accurate. The survival time O_i is the maximum survival time that the patient maintains in his or her state of existence at maximum observation time, the maximum observed time, that is, until the end of the final observation time, the patient is still alive. Time lost C_i refers to the first time a patient with a censored data appears to have censored data. If T_i is the observation time for the i patient, the dataset is valid when and only when the data set deletes the structure to the right and $T_i = \min(O_i, C_i)$ only. If you look at each patient as an instance in a survival analysis, you can use a ternary group (X_i, T_i, δ_i) to represent it. Where X_i is a $1 \times q$ feature vector, and q is the number of features (in this case, q is the number of days). δ_i is the censored data, which consists of 0 and 1. $\delta_i = 1$ represents the normal data, $\delta_i = 0$ represents the censored data, while the maximum observation time T_i is equal to the lifetime O_i or the censored time C_i, and the formula is as follows:

$$T_i = \begin{cases} O_i & \text{if} \quad \delta_i = 1 \\ C_i & \text{if} \quad \delta_i = 0. \end{cases} \tag{7}$$

Suppose that there are n patients in the survival data, that is n instances, then the feature matrix X is a $n \times q$ matrix. The purpose of establishing multi-task learning model is to use the ternary group (X_i, T_i, δ_i) to analyze the relationship between X_i and T_i in the survival data, where $i = 1, \ldots, n$.

But in the actual model, the observation time T_i of the real number can not be used to train the multi-task learning model. In this paper, we divide the observed time T_i into convenient integer time intervals, rather than using the observation time of the real number directly, and then transforming the time interval after partition into the target matrix Y, which is converted according to the censored indicator δ_i into two columns, and $k = \lfloor \max(T_i) \rfloor, i = 1, \ldots, n$. The element in the target matrix Y has only two values, and the "0" is used to represent the patient's death at i intervals, and "1" represents the patient's survival at i intervals. For the censored data, only the censored time is known from the survival data, and the patient's state of existence can not be obtained directly after censored time. At this point, the traditional survival analysis problem in this paper has been transformed into multi-task learning problem. The reason to use the multi-task learning model to solve the survival analysis is that the shared representation of related tasks in multi-task learning can accurately capture the dependencies between time intervals, the use of these dependencies will greatly reduce the predictive error of each task.

Therefore, in order to better use the censored data for model training, the target matrix Y uses "?" to represent the censored data, and then to further transform the target matrix Y. Fig. 1 shows how the target matrix Y is generated from the original survival data.

The matrix on the left is the original survival data, and on the right is the corresponding target matrix Y. For the following three normal data $ID = 1, 4, 5$, the patient is marked as "1" in the viewing time $(T_1 = 6, T_4 = 4, T_5 = 5)$, and the time in the state of death is marked as "0". In the case of censored data $ID = 2, 3$, the state before the censored time is marked as "1", the state after censored time is marked "?", which means that the data has been lost. In practice, consider the following:

(1) the patient will only go from being alive to dying and not being able to return to a state of being.
(2) once the patient's survival data begins to be lost, the patient's normal state (state of existence or state of death) can not be obtained if the patient's survival data begins to be lost.

Authorized licensed use limited to: South China Agricultural Univ. Downloaded on June 18,2021 at 08:21:57 UTC from IEEE Xplore. Restrictions apply.

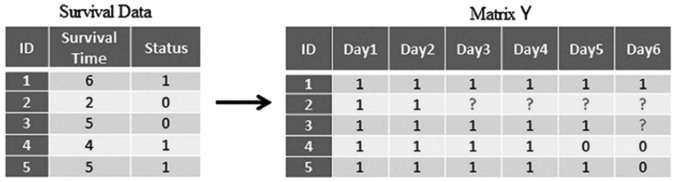


Fig. 1. A schematic diagram of the corresponding target matrix Y is generated from the original survival data.

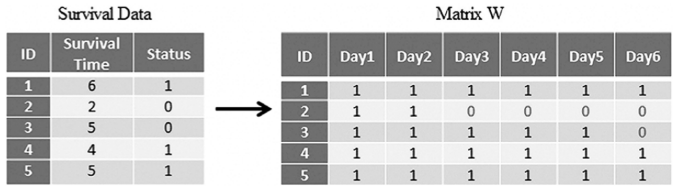


Fig. 2. A schematic diagram of the corresponding target matrix W is generated from the original survival data.

For this reason, we can make the following assumption:

(1) for each row in the target matrix Y, once the label has changed from "1" to something "0", it does not change from something "0" to "1", that is, death can not be revived.
(2) for each row in the target matrix Y, once the label "?" appears, until the maximum viewing time, the tag is "?".

Therefore, this paper defines the above-mentioned properties as non-negative non-incremental properties. The non-negative non-incremental properties of the data in the target matrix Y can be expressed in the following formula:

$$\begin{aligned} &P = \{Y \geq 0, Y_{ij} \geq Y_{il} | j \leq l, \\ &\forall j = 1, \ldots, k, \forall l = 1, \ldots, k\}, i = 1, \ldots, n. \end{aligned} \tag{8}$$

In the light of the above definition, the following multi-task learning formulae are given to solve the genetic analysis of cancer

$$\underset{XB \in P}{\text{minimize}} \frac{1}{2} \|Y - XB\|_F^2 + R(B). \tag{9}$$

In the formula, $B \in \mathbb{R}^{q \times k}$ is the estimation coefficient matrix, the $\|\cdot\|_F$ represents the Frobenius norm, and the $R(B)$ is an artificial penalty term to prevent over-fitting which includes additional constraints on the formula. However, the matrix Y is not a complete $n \times k$ target matrix (there is censored data in the survival data, and the target matrix Y contains "?" items) . Therefore, a directive matrix W is defined to deal with the problem of the presence of "?" in the matrix Y. The matrix W is a $n \times k$ two-dimensional directive matrix. If the state of existence in the survival data is deterministic, that is, $Y_{ij} = 1$ or $Y_{ij} = 0$, then $W_{ij} = 1$. If the state of existence in the survival data is uncertain, that is $Y_{ij} = ?$, then $W_{ij} = 0$. Fig. 2 shows how to generate the indicator matrix W from the original survival data.

Therefore, the multi-task learning problem defined in Equation (5) can be rewritten in the following form

$$\begin{aligned} &\underset{XB \in P}{\text{minimize}} \frac{1}{2} \left\| \prod\nolimits_W (Y - XB) \right\|_F^2 + R(B) \\ &\prod\nolimits_W U_{ij} = \begin{cases} U_{ij} & if \quad W_{ij} = 1 \\ 0 & if \quad W_{ij} = 0. \end{cases} \end{aligned} \tag{10}$$

In the experiment of this paper, the multi-task learning algorithm can not handle the text data of "?" . In order to better apply the censored data to the multi-task learning model and keep the data non-negative and non-incremental, the real "0.5" is used instead of "?" for model training, so that the data can be better used to reduce the error. In analyzing the survival data, we find that it has two unique properties: the state of patient is non-negative and non-recursive, time-smooth. Besides the above two properties, Frobenius norm penalty item is added to the coefficient matrix B, the problem of overfitting of high-dimensional data can be alleviated well. And we chose the $L_{2,1}$ norm as another norm penalty, because the $L_{2,1}$ norm allows multiple predictors to share similar sparse patterns. At the same time, the $L_{2,1}$ norm penalty can pick out some common features in all tasks, and this sparseness can effectively help the model process the galway dataset. The two properties of the above definition and the norm penalty are added to the multi-task learning model, transforming the Formula (10) into the following formulae of minimization:

$$\underset{XB \in P}{\text{minimize}} \frac{1}{2} \left\| \prod_W (Y - XB) \right\|_F^2 + \frac{\lambda_1}{2} \|B\|_F^2 + \lambda_2 \|B\|_{2,1}. \tag{11}$$

In the formula, $\|\cdot\|_{2,1}$ is a $L_{2,1}$ norm penalty term, λ_1 and λ_2 as parameters for two penalty items, $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$.

However, this optimization is not enough because the optimization still contains non-negative non-incremental constraints, and the introduced $L_{2,1}$ norm penalty does not have a time smoothness. The ADMM algorithm can also be used here to solve the optimization problem presented in Formula (11) .

ADMM algorithm in order to increase the robustness of the dual-ascending method and to minimize the strong constraints of the function, we need to introduce the extended lagrange form. A new matrix $M = XB$ is defined, and the Formula (11) is overridden as an ADMM algorithm in the following form:

$$\begin{aligned} \underset{M \in P}{\text{minimize}} &\frac{1}{2} \left\| \prod\nolimits_W (Y - M) \right\|_F^2 + \frac{\lambda_1}{2} \|B\|_F^2 \\ &+ \lambda_2 \|B\|_{2,1} \\ &\text{s.t.} M = XB. \end{aligned} \tag{12}$$

Then the residual term μ and the penalty coefficient ρ were added to the model, of which $\rho > 0$. The increase in Formula (12) is obtained in the form of Joseph-Louis Lagrange

$$\begin{aligned} L_\rho(M, B, \mu) = &\frac{1}{2} \left\| \prod\nolimits_W (Y - M) \right\|_F^2 + \frac{\lambda_1}{2} \|B\|_F^2 \\ &+ \lambda_2 \|B\|_{2,1} + \frac{\rho}{2} \|M - XB + \mu\|_F^2. \end{aligned} \tag{13}$$

In Formula (13), $\frac{\rho}{2} \|M - XB + \mu\|_F^2$ is an additional penalty. Thus, the scaling form of the Formula (13) can be written as

Authorized licensed use limited to: South China Agricultural Univ. Downloaded on June 18,2021 at 08:21:57 UTC from IEEE Xplore. Restrictions apply.

$$M^{t+1} := \arg\min_{M \in P} \left(\frac{1}{2} \left\| \prod_W (Y - M) \right\|_F^2 + \frac{\rho}{2} \| M - XB^t + \mu^t \|_F^2 \right) \tag{14}$$

$$B^{t+1} := \arg\min_{B \in \mathbb{R}^{q \times k}} \left(\frac{\lambda_1}{2} \|B\|_F^2 + \lambda_2 \|B\|_{2,1} + \frac{\rho}{2} \| M^{t+1} - XB + \mu^t \|_F^2 \right) \tag{15}$$

$$\mu^{t+1} := \mu^t + M^{t+1} - XB^{t+1}. \tag{16}$$

The calculation of the ADMM algorithm is a repeated iterative Formula (14) and (15) , and then uses the residual Formula (16) to control the number of iterations. When the precision reaches the accuracy of the model, the iteration is stopped and the optimal matrix X and matrix C are obtained. The process of solving the optimization problem in the Formula (14) and (15) is described in detail in steps below.

Step 1: Update the M^{t+1} with a given B^t and μ^t.

The update of M^{t+1} (Formula (14)) is a smooth convex optimization problem that can be expressed in the following broad form:

$$\begin{aligned} &\min_{M \in P} g(M) \\ &g(M) = \frac{1}{2} \left\| \prod_W (Y - M) \right\|_F^2 + \frac{\rho}{2} \left\| M - XB^t + \mu^t \right\|_F^2. \end{aligned} \tag{17}$$

Because $g(M)$ is a smooth convex function with nonnegative non-incremental properties with the above definition. Suppose $S = \mu^t - XB^t$, and then decompose the function $g(M)$

$$\begin{aligned} g(M) &= \frac{1}{2} \left\| \prod_W (Y - M) \right\|_F^2 + \frac{\rho}{2} \| M - XB^t + \mu^t \|_F^2 \\ &= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{k} W_{ij} (Y_{ij} - M_{ij})^2 + \frac{\rho}{2} \sum_{i=1}^{n} \sum_{j=1}^{k} (M_{ij} + S_{ij})^2 \\ &= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{k} \left[\begin{array}{c} (W_{ij} + \rho) M_{ij}^2 + 2(\rho S_{ij} - Y_{ij} W_{ij}) M_{ij} \\ + W_{ij} Y_{ij}^2 + \rho S_{ij}^2 \end{array} \right] \\ &= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{k} (W_{ij} + \rho) \left[\begin{array}{c} M_{ij}^2 + \frac{2(\rho S_{ij} - Y_{ij} W_{ij}) M_{ij}}{W_{ij} + \rho} + \\ \left(\frac{\rho S_{ij} - Y_{ij} W_{ij}}{W_{ij} + \rho} \right)^2 - \left(\frac{\rho S_{ij} - Y_{ij} W_{ij}}{W_{ij} + \rho} \right)^2 \end{array} \right] \\ &\quad + \frac{W_{ij} Y_{ij}^2 + \rho S_{ij}^2}{W_{ij} + \rho} \\ &= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{k} (W_{ij} + \rho) \left[\left(M_{ij} - \frac{Y_{ij} W_{ij} - \rho S_{ij}}{W_{ij} + \rho} \right)^2 + Q_{ij} \right]. \end{aligned} \tag{18}$$

In the equation, $Q_{ij} = \frac{W_{ij} Y_{ij}^2 + \rho S_{ij}^2}{W_{ij} + \rho} - (\frac{\rho S_{ij} - Y_{ij} W_{ij}}{W_{ij} + \rho})^2$. The formula shows that Q_{ij} does not depend on M_{ij}. Thus, the convex optimization problem in Formula (17) is equivalent to the Euclid projection problem. The following formulae can be obtained by removing all extraneous items from M

$$M^{t+1} = \min_{M \in P} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{k} (M_{ij} - \widetilde{M}_{ij})^2. \tag{19}$$

$\widetilde{M}_{ij} = \frac{Y_{ij} W_{ij} - \rho S_{ij}}{W_{ij} + \rho}$ needs to be projected onto the set P and ensure $M_{i1} \geq M_{i2} \geq \cdots \geq M_{ik} \geq 0$. The Euclid projection in Formula (18) is a special case that requires the use of non-negative maximum heap projections to effectively solve the problem.

Step 2: Update B^{t+1} with a given M^{t+1} and μ^t

The update of B^{t+1} (Formula (15)) is a standard $L_{2,1}$ normalization problem that can be converted into the following forms:

$$\arg\min_{B \in \mathbb{R}^{q \times k}} \left(\frac{1}{2} \|L - XB\|_F^2 + \rho_{L2} \|B\|_F^2 + \rho_{L1} \|B\|_{2,1} \right). \tag{20}$$

In the formula, L is the corresponding label matrix. In order to solve the Formula (15), set $L = M^{t+1} + \mu^t$, $\rho_{L1} = \frac{\lambda_2}{\rho}$, $\rho_{L2} = \frac{\lambda_1}{2\rho}$, and then use alternating minimization or Nesterov Algorithm to solve the Euclid problem in Formula (15) . In this paper, we use the Nesterov Algorithm [21], [30] which is used to solve the $L_{2,1}$ norm. The Algorithm needs $O(\frac{1}{\sqrt{\tau}})$ iterations to achieve the accuracy of τ.

Combining all of the above, the algorithm steps of the multi-task learning model can be summarized as follows. The M^0 is initialized as a target matrix, in which $M^0 = Y$ is repeated, and the two steps are iterated to update the M^t and B^t until the accuracy required for the model is achieved. The pseudo code for the multi-task learning model presented in this article is shown in Fig. 3.

```
Input: Feature Matrix X, Target Matrix Y, indicative Matrix W, ρ, λ1, λ2
Output: B̂
1. Initialization: t = 0, M^t = Y, μ^t = 0, B^t = 0
2. The Cycle Begins
3.     Calculate M^{t+1} by using B^t and μ^t
4.     Set L = M^{t+1} + μ^t, ρ_{L1} = λ2/ρ, ρ_{L2} = λ1/2ρ
5.     Calculate B^{t+1} by using M^{t+1} and μ^t
6.     Calculate μ^{t+1} = μ^t + M^{t+1} − XB^{t+1}
7.     t = t + 1
8. End when it converges to the desired precision.
9. B̂ = B^t
```

Fig. 3. The pseudo code for the multi-task learning model.

4.2 Analysis of Multi-Task Learning Algorithm

4.2.1 Convergence Analysis

The Formula (12) is a standard multi-task learning formula. The standard ADMM can be expressed as the following formula:

$$\underset{M \in P, B \in \mathbb{R}^{q \times k}}{\text{mininmize}} f_1(M) + f_2(B) \ s.t. \ M = XB. \tag{21}$$

In the formula, $f_1(M) = \frac{1}{2} \| \prod_W (Y - M) \|_F^2$

$f_2(B) = \frac{\lambda_1}{2} \|B\|_F^2 + \lambda_2 \|B\|_{2,1}$. Besides, $f_1(M)$ and $f_2(B)$ are convex functions, and P and $\mathbb{R}^{q \times k}$ are closed convex sets.

According to the definition of the ADMM algorithm, only the $k+1$ iteration of the residual error γ is less than the precision τ. That is to say when $\gamma^{k+1} = Ax^{k+1} + Bz^{k+1} - c < \tau^{primal}$, the algorithm must converge. According to the proof of He Bingsheng et al. [14], if the precision is τ, the algorithm of ADMM only needs to iterate to $O(\frac{1}{\tau})$ times, the algorithm is basically convergent. In the experiment of this paper, an iteration termination condition with precision of $\tau < 1 \times 10^{-4}$ is set to ensure more accurate conclusion.

Authorized licensed use limited to: South China Agricultural Univ. Downloaded on June 18,2021 at 08:21:57 UTC from IEEE Xplore. Restrictions apply.

TABLE 1
Detail of the Dataset Used in this Article

Dataset	Instances	Features	Censored	Tasks
NSBCD	115	549	77	188
VDV	78	4705	44	13
AML	116	6283	49	54
Lung	86	7129	62	110
MCL	92	8810	28	14
DBCD	295	4919	216	18

4.2.2 Time Complexity Analysis

The multi-task learning model uses the ADMM algorithm to solve the CONVEX function problem optimally. The main steps are divided into two steps-the update of M^{t+1} and the update of B^{t+1}.

First, calculating the time complexity of the Formula (18). The time complexity of $\widetilde{M}_{ij}$ in Formula (18) is $O(q)$ because only S_{ij} in $\widetilde{M}_{ij}$ has a matrix multiplication, while X_i in $S_{ij} = \mu_{ij}^t - X_i B_{(,j)}^t$ (the i row of the eigenmatrix) and $B_{(,j)}^t$ (row j in B^t) both have q elements. In the update of M^{t+1}, M is $n \times k$ matrix. It needs time complexity $O(nqk)$ to calculate $n \times k$ $\widetilde{M}_{ij}$. The time complexity of updating M^{t+1} is $O(nqk)$, where n, q and k denote the number of instances, feature dimensions and tasks, respectively.

Then calculate the time complexity of the Formula (19). According to the conclusion in the previous section, because the B^{t+1} update addresses the least squares of regularizing the $L_{2,1}$ norm, the accuracy of the entire iteration reaches τ, which requires time complexity $O(\frac{1}{\sqrt{\tau}}nqk)$.

Therefore, based on the convergence rate of multi-task learning algorithm, the total time complexity of the model is $O(\frac{1}{\tau\sqrt{\tau}}nqk)$, which can reach the convergence accuracy of τ^{primal} class.

5 Experiments

5.1 Source of Data

The survival data used in this model are derived from several publicly available high dimensional gene expression cancer datasets, including data on gene expression in diseases such as breast, lung, leukemia, and cell lymphoma, details are as follows:

(1) Norwegian Stanford's breast cancer data (NSBCD). They included gene expression sequencing in 115 women with breast cancer, and medical institutions continued to observe the women for 188 months and record their survival.
(2) Van de Vijver's microarray breast cancer data (VDV). It contains information about gene expression that can be used to predict the survival of breast cancer. The dataset contains 13 years of survival information for 78 patients and contains a total of 4,707 gene expressions.
(3) adult myeloid leukemia data (AML). They included 116 AML patients with Gene expression profiling with a maximum follow-up of 1,625 days. In the experiment, the unit of observation time could be changed from daily to monthly and observed for 54 months.
(4) Lung Cancer (Lung) is a Gene expression profiling of 86 cases of early Lung cancer with a duration of 110 months.
(5) neoplastic lymphoma (MCL) is a 14-year survival data collected from 92 MCL patients.
(6) Dutch breast cancer data (DBCD). Survival information on the expression of 4,919 genes in 29,519 women with breast cancer was 18 years the maximum.

The data in the six open datasets are described in detail in Table 1. The 'Instances' column and the 'Features' column respectively represent the number of instances corresponding to the data set and the number of characteristics corresponding to each instance. The 'Censored' is listed as the number of instances in which data is censored. In the survival prediction of cancer genes, the event of interest is when the patient dies. Thus, the uncensored cases represent patients who died during the observation period, while the censored cases represent patients who remain alive until the time was censored. According to the observation time of each data set, the problem of survival prediction can be transformed into the corresponding multi-task learning problem. The number of tasks for each data is given in the 'Tasks' column in the table, as well as representing the number of features. Because there is a need to distinguish between training sets and test sets, different cross-validation is used for different instances. When the number of instances is greater than 150, a 5-fold cross-validation is used and a 3-fold cross-validation is used when the number of instances is less than 150.

5.2 The Method of Computing

Several mainstream related algorithms are compared with the algorithms used in this paper as well as the traditional methods. We classify the algorithms and then introduce the method of computing for all the used models in this paper.

(1) regression model of parameter review, including Logistic and Weibull.
(2) linear regression model, including Tobit and OLS.
(3) Cox proportional risk regression model, including COX, LASSO-COX and EN-COX.
(4) multi-task learning algorithm, including IMACS (The method proposed in this paper) and traditional multi-task learning algorithm.

5.2.1 Regression Model of Parameter Review

We can use the product of the death density function to express the joint probability of the unexamined instances, and use the product of the survival function to express the joint probability of the examined instances. Then the combination of these two parts gives us a standard likelihood function, whose model parameters can be estimated by using the maximum likelihood estimation procedure [19].

5.2.2 Linear Regression Model

For a linear regression function $y_j = X_j\beta + \varepsilon_j, \varepsilon_j \sim N(0, \sigma^2)$, Tobin proposed an extension of it called Tobit model [28]. But he uses maximum likelihood to estimate parameters instead of least squares errors. Tobit model can be trained by the survreg function containing Gaussian distributions. Besides, ordinary least squares (OLS) has a loss function that

Authorized licensed use limited to: South China Agricultural Univ. Downloaded on June 18,2021 at 08:21:57 UTC from IEEE Xplore. Restrictions apply.

TABLE 2
Comparing the Multitask Learning Model with Other Models by Using the Value of C-Index

Model	NSBCD	VDV	AML	Lung	MCL	DBCD
Logistic	0.2795	0.4446	0.4357	0.6643	0.5087	0.4960
Weibull	0.5335	0.2329	0.5245	0.3955	0.2949	0.5599
Tobit	0.4441	0.6666	0.5076	0.4966	0.4611	0.4820
OLS	0.6246	0.5990	0.5031	0.4957	0.5848	0.5429
COX	0.3963	0.6025	0.5434	0.3738	0.6254	0.5363
LASSO-COX	0.5423	0.5993	0.4922	0.5856	0.6981	0.7292
EN-COX	0.5567	0.6058	0.6169	**0.6898**	0.7064	**0.7705**
Traditional MTL	0.5624	0.6190	0.4103	0.4822	0.6976	0.6092
IMACS	**0.6973**	**0.7215**	**0.7187**	0.6624	**0.7485**	0.7557

similar to the one in our model. We also use it to compare, although it is actually not a censored regression model. And we only use uncensored data to train it instead of entire set.

5.2.3 Cox Proportional Risk Regression Model

The hazard function of Cox proportional risk regression model is $\lambda(\mathrm{t}, X_i) = \lambda_0(t) \cdot \exp(X_i\beta)$, where $\lambda_0(t)$ is the common baseline hazard function for all instances, and β is a coefficient vector estimated by minimizing the negative logarithmic partial likelihood function. We can use the coxph function in the survival package to train the Cox model [27]. And we can use the fastcox package to train LASSO-COX and EN-COX [31].

5.2.4 Multi-Task Learning Algorithm

The traditional multi-task learning algorithm can only use the uncensored data in the survival data to train the model, the model's generalization ability has been reduced.

5.3 The Performance of the Algorithm

Because of the existence of censored data in the survival data, the standard evaluation indicators used for regression analysis (for example, sum of the mean square errors R^2) can not be used for the performance comparison of the model above, the consistency index and weighted average AUC often appear in the field of survivability analysis, which is used to measure the performance index of the prediction model.

5.3.1 Concordance Index (C-Index)

C-index was first proposed by Frank e Harrell Jr, a professor of biostatistics at Vanderbilt University, in 1996, to evaluate the distinction between real and predicted values in a survival analysis, the prediction accuracy of the prognosis model of cancer cancer patients was used more frequently. In clinical application, the precision of prediction is usually emphasized, and the C-index is often used as the prediction precision in the model evaluation. In survival analysis, C-index is to group all the research instances into a binary set, each with its own predictive and real values. If two of the patients in the binary group lived longer than the other if they had a longer survival time, or the one with a higher probability of survival, this is called the prediction result, which is consistent with the actual result.

In this paper, $(y_1, \hat{y}_1)$ and $(y_2, \hat{y}_2)$ of a binary variable group are considered, where y_i is the actual observation value and $\hat{y}_i$ is the predicted value. The probability of consistency can be expressed as follows:

$$c = Pr(\hat{y}_1 > \hat{y}_2 | y_1 > y_2). \tag{22}$$

According to the definition of C-index, it is found that the area (AUC) under the ROC curve in the binary classification has the same proportion with C-index. If y_i is binary, then the value of a C-index is equal to the value of AUC. According to the definition in the Cox model, instances with lower risk rates can survive longer, so the value of the C-index will be calculated in the following formula:

$$c = \frac{1}{\mathrm{num}} \sum_{i \in \{1,\ldots,N\}, \delta_i = 1} \sum_{y_j > y_i} I(X_i\hat{\beta} > X_j\hat{\beta}). \tag{23}$$

The "num" represents the number of $(y_i, \hat{y}_i)$ in the binary group, the $I[\cdot]$ as the index function, the "1" representing the uncensored data, and the "0" representing the censored data. In other algorithms, the value of C-index is expressed as

$$c = \frac{1}{\mathrm{num}} \sum_{i \in \{1,\ldots,N\}, \delta_i = 1} \sum_{y_j > y_i} I[S(\hat{y}_j | X_j.) > S(\hat{y}_i | X_i.)]. \tag{24}$$

$S(\hat{y}_i|X_i.)$ is the predicted target value. Multi-task learning models can not predict survival time directly, but can determine whether an instance is in a state of existence at each time interval. So it's possible to predict survival time indirectly. In the experiment, the number of binary groups was not counted because for the case containing censored data, the maximum review time of it is actually censored time, which would have resulted in a greater error if used directly to calculate the value of C-index, so the formula uses the index function to filter out instances where the data is uncensored to calculate the value of C-index.

The Table 2 compares the multitask learning model (IMACS) with other models by using the value of C-index. In the table, several open high-dimensional gene expression cancer survival data sets are trained using the multi-task learning model and other relevant models in this paper. After ten training sessions on various data sets, we get the average value of C-index. The best data in the table will be highlighted. The experimental results show that the improved multi-task learning algorithm used in this paper has good performance in all six data sets, accuracy is 0.7 and above, and is basically higher than other models. Therefore, the multi-task learning

Authorized licensed use limited to: South China Agricultural Univ. Downloaded on June 18,2021 at 08:21:57 UTC from IEEE Xplore. Restrictions apply.

TABLE 3
Comparing the Multitask Learning Model with Other Models by Using Weighted Average AUC Value

Model	NSBCD	VDV	AML	Lung	MCL	DBCD
Logistic	0.2745	0.5547	0.3986	0.6236	0.5658	0.5578
Weibull	0.4709	0.2683	0.6037	0.5034	0.3665	0.5491
Tobit	0.4740	0.6636	0.4928	0.5355	0.4319	0.4530
OLS	0.6738	0.5231	0.5040	0.6346	0.4899	0.5200
COX	0.4322	0.5582	0.6436	0.4318	0.4878	0.6406
LASSO-COX	0.5914	0.5606	0.5257	0.6231	0.6156	0.7854
EN-COX	0.6004	0.6964	0.4792	0.7078	0.6838	**0.8168**
Traditional MTL	0.5766	0.4499	0.5131	0.3595	0.7177	0.6997
IMACS	**0.6966**	**0.7159**	**0.6438**	**0.7108**	**0.7533**	0.7819

model proposed in this paper is superior to other models in predicting performance.

5.3.2 Weighted Average AUC

The value of C-index can be used to assess the performance of the model in the regression problem. In addition, the performance of the model in classification problems needs to be evaluated, and the classification problem in the survival analysis divides the patients into two categories, namely survival and death. The number of patients with a known state of existence (target matrix Y labelled as "1" or "0") will decrease over time due to the presence of censored data. In Table 3, the weighted average AUC values with different time intervals were used for comparison. The weighted average AUC value is defined as follows:

$$\mathrm{AUC}_{\mathrm{avg}} = \frac{\sum_{i=1}^{k} \mathrm{AUC}^{(i)} n_{\bar{c}}^{(i)}}{\sum_{i=1}^{k} n_{\bar{c}}^{(i)}}. \tag{25}$$

The $\mathrm{AUC}^{(i)}$ is the AUC value corresponding to the task i, and $n_{\bar{c}}^{(i)}$ is the number of instances with known survival status tags within the i interval.

The Table 3 compares the multitask learning model (IMACS) with other models by using weighted average AUC value. The experimental results show that the model can get higher $\mathrm{AUC}_{\mathrm{avg}}$ value in most data sets, and the AUC is 0.7 and above, and the representative model can predict the survival state of the patient at every time interval, this suggests that the improved models used have the ability to make better use of time intervals to predict than other models.

5.4 Model Scalability Test

In this section, we analyze the scalability of survival analysis based on multi-task learning model. At the end of the experiment, the total time complexity of the model was $O(\frac{1}{\tau\sqrt{\tau}} nqk)$, where n was an instance quantity, q was the characteristic quantity, and k was the maximum observation time. In order to test the scalability of the model proposed in this paper, a random set of data is generated using 'simple.surv.sim' in the survsim package in R. The generated datasets have different number of instances, different characteristic quantities and different maximum observation time. When studying the relation between the number of cases n and the running time, we need to fix the characteristic number q and the maximum observation time k, change the number of cases n, and run the multi-task learning model 100 times to get the time-consuming Diagram (Fig. 4).

When studying the relation between the characteristic number q and the running time, we need to fix the number of cases n and the maximum observation time k, change the characteristic number q, and run the multi-task learning model 100 times to get the time-consuming diagram (Fig. 5).

When studying the relation between the maximum observation time k and the running time, we need to fix the number of cases n and the characteristic number q, change the maximum observation time k, and run the multi-task learning model 100 times to get the time-consuming diagram (Fig. 6).

All of these randomly generated data are uniformly distributed, with random intervals between each of the different features, and the coefficient matrix B will also randomly generate the range of each element in the $[-1, 1]$ range. The

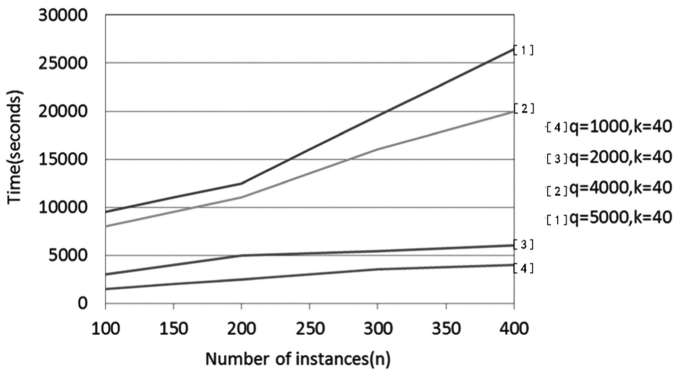


Fig. 4. Time-consuming diagram of run the multi-task learning model 100 times (change the number of cases n).

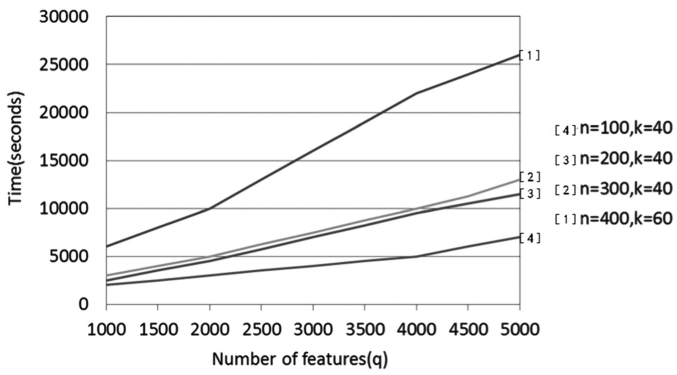


Fig. 5. Time-consuming diagram of run the multi-task learning model 100 times (change the characteristic number q.).

Authorized licensed use limited to: South China Agricultural Univ. Downloaded on June 18,2021 at 08:21:57 UTC from IEEE Xplore. Restrictions apply.

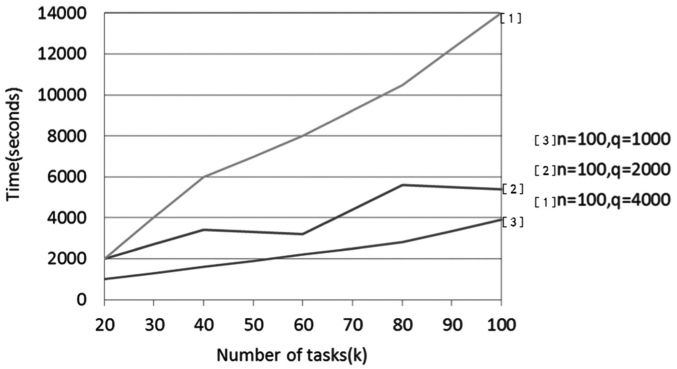


Fig. 6. Time-consuming diagram of run the multi-task learning model 100 times (change the maximum observation time k.).

random survival time O_i follows the Log-logistic distribution and the time C_i obeys the Weibull distribution, which is the training standard for survival analysis. From the above three pictures, we can see that the operation time of the multi-task learning model is linear relative to the number of instances n, maximum observation time k, and characteristic number q, the results show that the experiment has good stability in the slow growth of the data.

6 Conclusion

This paper proposes an improved muti-task learning algorithm for analyzing cancer survival data, which defines the original problem of survival time as multi-task learning problem, it breaks through the limitation that other algorithms cannot handle censored data or need to be strictly assumed. It greatly enhances the practicability and generality of the algorithm, and improves the prediction accuracy by $10 \sim 15$ percent compared with other mainstream algorithms.

Using the method proposed in this paper, medical institutions can better predict the patient's survival time and improve the efficiency of comprehensive clinical examination. The results obtained by this method are credible and can provide a reliable reference for clinical medicine, to help doctors guide patients and assess whether treatment is appropriate for patients. The stability of the method is further enhanced to facilitate the determination of the drug effect and to speed up drug development.

Based on this article, the research direction also has the following worthy promotion:

1) Although this paper used multi-task learning model, which processed the censored data and trained. However, it is still not possible to predict censored data by means of a survival analysis. In this paper, we only use the existence state of the censored data which has not reached the censored time, and the unknown state after the censored time cannot be used in the model. In the next study, further analysis and processing of the missing data will be carried out to enable the model to make better use of all the censored data.
2) When using ADMM algorithm to optimize the multi-task learning formula, the number and precision of iteration used in this paper refer to the standard proposed by Bingsheng H et al [14]. Because the accuracy of different models is different, in the next study, we need to do more research on accuracy to get more accurate predictions.

Acknowledgments

This work was financially supported by the Guangdong Natural Science Foundation Project (2018A030313437), the Ministry of Education Humanities and Social Sciences Research Youth Fund Project (18YJCZH037), and the Guangdong Science and Technology Program Project (2018A070712021).

References

[1] Z. Y. Algamal, M. H. Lee, A. M. Al-Fakih, and M. Aziz, "High-dimensional QSAR modelling using penalized linear regression model with L1/2-norm," *Sar Qsar Environ. Res.*, vol. 27, no. 9, pp. 703–719, 2016.
[2] B. Steffen, J. Bogojeska, T. Lengauer, T. Scheffer, "Multi-task learning for HIV therapy screening," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 56–63.
[3] S. Boyd, N. Parikh, E. Chu, and B. Peleato, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2010.
[4] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.
[5] D. R. Cox, "Regression models and life-tables," *J. Roy. Statistical Soc.*, vol. 34, no. 2, pp. 187–220, 1972.
[6] T. Evgeniou and M. Pontil, "Regularized multi-task learning," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 109–117.
[7] H. Fei and J. Huan, "Structured feature selection and task relationship inference for multi-task learning," in *Proc. IEEE Int. Conf. Data Mining*, 2012, pp. 171–180.
[8] S. Fotso, "Deep neural networks for survival analysis based on a multi-task framework," *arXiv preprint arXiv:1801.05512*, 2018.
[9] T. A. Gerds and M. Vogeler, "Endpoints and survival analysis for successful osseointegration of dental implants," *Statistical Methods Med. Res.*, vol. 14, no. 6, pp. 579–590, 2005.
[10] A. Guilloux, S. Lemler, and M. L. Taupin, "Adaptive estimation of the baseline hazard function in the Cox model by model selection, with high-dimensional covariates," *J. Statistical Planning Inference*, vol. 148, no. 2, pp. 141–159, 2016.
[11] Y. Guo and W. Xue, "Probabilistic multi-label classification with sparse feature learning," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 1373–1379.
[12] B. N. Harris, A. A. Bhuskute, S. Rao, D. G. Farwell, and A. F. Bewley, "Primary surgery for advanced-stage laryngeal cancer: A stage and subsite-specific survival analysis," *Head Neck*, vol. 38, no. 9, pp. 1380–1386, 2016.
[13] M. R. Hassan, M. A. Suan, S. A. Soelar, N. S. Mohammed, I. Ismail, and F. Ahmad, "Survival analysis and prognostic factors for colorectal cancer patients in malaysia," *Asian Pacific J. Cancer Prevention*, vol. 17, no. 7, 2016, Art. no. 3575.
[14] B. He and X. Yuan, "On the O(1/n) convergence rate of the Douglas-Rachford alternating direction method," *SIAM J. Numerical Anal.*, vol. 50, no. 2, pp. 700–709, 2012.
[15] D. E. Ho, K. Imai, G. King, and E. A. Stuart, "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference," *Political Anal.*, vol. 15, pp. 199–236, 2007.
[16] J. L. Horowitz and G. R. Neumann, "Specification testing in censored regression models: Parametric and semiparametric methods," *J. Appl. Econometrics*, vol. 4, no. S1, pp. S61–S86, 2010.
[17] J. Jankova and S. Van De Geer, "Semi-parametric efficiency bounds for high-dimensional models," *Ann. Statist.*, vol. 46, no. 5, pp. 2336–2359, 2018.
[18] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Publications Amer. Statistical Assoc.*, vol. 53, no. 282, pp. 457–481, 1958.
[19] E. T. Lee and J. Wang, *Statistical Methods for Survival Data Analysis*. Hoboken, NJ, USA: Wiley, 2003.
[20] Y. Li, K. S. Xu, and C. K. Reddy, "Regularized parametric regression for high-dimensional survival analysis," in *Proc. SIAM Int. Conf. Data Mining*, 2016, pp. 765–773.

Authorized licensed use limited to: South China Agricultural Univ. Downloaded on June 18,2021 at 08:21:57 UTC from IEEE Xplore. Restrictions apply.

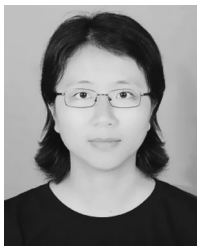
[21] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient l2, 1-Norm minimization," in *Proc. 25th Conf. Uncertainty Artif. Intell.*, AUAI Press, 2009, pp. 339–348.
[22] M. Morenobetancur, H. Sadaoui, C. Piffaretti, and G. Rey, "Survival analysis with multiple causes of death: Extending the competing risks model," *Epidemiology*, vol. 28, no. 1, 2017, Art. no. 12.
[23] A. A. Mostafa and A. B. Ghorbal, "Using WinBUGS to cox model with changing from the baseline hazard function," *Social Sci. Electron. Publishing*, vol. 5, pp. 2217–2240, 2015.
[24] S. Özdemir and O. Arslan, "An alternative algorithm of the empirical likelihood estimation for the parameter of a linear regression model," *Commun. Statist.- Simul. Comput.*, vol. 502, pp. 1–9, 2018.
[25] Y. Qi, O. Tastan, J. G. Carbonell, J. Klein-Seetharaman, and J. Weston, "Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins," *Bioinf.*, vol. 26, no. 18, pp. i645–i652, 2010.
[26] M. A. Rubeo and T. L. Schmitz, "Mechanistic force model coefficients: A comparison of linear regression and nonlinear optimization," *Precision Eng.*, vol. 45, pp. 311–321, 2016.
[27] T. Therneau, "A package for survival analysis in S. R package version 2.37–4," 2013. [Online]. Available: http://CRAN.R-project.org/package=survival
[28] J. Tobin, "Estimation of relationships for limited dependent variables," *Econometrica*, vol. 26, no. 1, pp. 24–36, 1958.
[29] W. Lu, Y. Li, J. Zhou, D. Zhu, and J. Ye, "Multi-task survival analysis," in *Proc. IEEE Int. Conf. Data Mining*, 2017, pp. 485–494.
[30] S. Wang, B. Nan, J. Zhu, and D. G. Beer, "Doubly penalized buckley-james method for survival data with highdimensional covariates," *Biometrics*, vol. 64, no. 1, pp. 132–140, 2010.
[31] Y. Yang and H. Zou, "A cocktail algorithm for solving the elastic net penalized cox's regression in high dimensions," *Statist. Interface*, vol. 6, no. 2, pp. 167–173, 2012.
[32] W. Zhang, R. Li, T. Zeng, Q. Sun, S. Kumar, J. Ye, and S. Ji, "Deep model based transfer and multi-task learning for biological image analysis," *IEEE Trans. Big Data*, to be published, doi: 10.1109/TBDATA.2016.2573280.
[33] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multitask learning formulation for predicting disease progression," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 814–822.



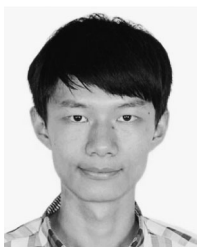
Wanrong Gu received the PhD degree in computer science and technology from the South China University of Technology, in 2015. Currently, he is an assistant professor with the Computer Science Department, South China Agricultural University. His main research interests include big data mining and forecasting, machine learning, and big data analysis of bioinformatics. He has published a number of papers in these fields.



Ziye Zhang is working toward the PhD degree in the College of Mathematical, South China University of Technology. His main research interests include mathematical modeling, big data analysis of bioinformatics, and machine learning. He has participated in a number of mathematical contest in modeling and won awards.



Xianfen Xie received the PhD degree in statistics from Jinan University, in 2016. Currently, she is an associate professor with the Statistics Department, Jinan University. Her main research interests include financial risk forecasting, big data analysis, and policy decision in economics. She has more than 10 years of research experience in these fields and has led a number of research projects in these fields.



Yichen He is working toward the graduate degree in the College of Mathematics and Informatics, South China Agricultural University. His main research interests include artificial intelligence, big data mining and forecasting, and machine learning. He has done in-depth research in these fields, participated in many related software development projects, and owned many software patents.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.

Authorized licensed use limited to: South China Agricultural Univ. Downloaded on June 18,2021 at 08:21:57 UTC from IEEE Xplore. Restrictions apply.

ISSN 1000-565X
CODEN HLDKEZ

華南理工大學學報

JOURNAL OF SOUTH CHINA UNIVERSITY OF TECHNOLOGY

自然科学版

Natural Science Edition

第49卷
1
2021年1月
Vol.49 No.1 (January 2021)
中国·广州 GUANGZHOU CHINA

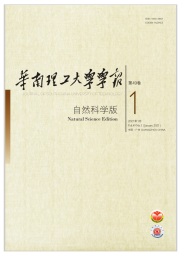
过刊浏览

(https://zrb.bjb.scut.edu.cn/CN/article/showOldVolumn.do)

2021年 第49卷 第1期　刊出日期：2021-01-25

上一期 (../volumn/volumn_1369.shtm
(../volumn/volumn_1

目次 (https://zrb.bjb.scut.edu.cn/CN/Y2021/V49/I1/0)



(https://zrb.bjb.scut.edu.cn/fileup/1000-565X/PIC/2021-1ml_pic.jpg)

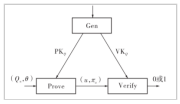
2021, 49(1): 0.

摘要 (43) PDF (408KB) (48)

相关文章 (https://zrb.bjb.scut.edu.cn/CN/Y2021/V49/I1/0#RelatedCitationTab) | 多维度评价 (https://zrb.bjb.scut.edu.cn/CN/Y2021/V49/I1/0#MetricsTab)

计算机科学与技术

一种基于 QAP 问题的 ZK-SNARK 新协议 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200207)



(https://zrb.bjb.scut.edu.cn/fileup/1000-565X/PIC/2020-0207_pic.jpg)

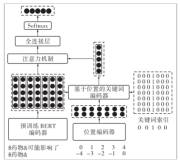
黄平 梁伟洁

2021, 49(1): 1-9. doi:10.12141/j.issn.1000-565X.200207 (https://doi.org/10.12141/j.issn.1000-565X.200207)

摘要 (360) PDF (685KB) (98)

相关文章 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200207#RelatedCitationTab) | 多维度评价 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200207#MetricsTab)

结合类别关键词与注意力机制的药物相互关系抽取模型 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200506)



(https://zrb.bjb.scut.edu.cn/fileup/1000-565X/PIC/2020-0506_pic.jpg)

IKA Novita Dewi, 蔡晓玲, 刘晓锋, 等

2021, 49(1): 10-17. doi:10.12141/j.issn.1000-565X.200506 (https://doi.org/10.12141/j.issn.1000-565X.200506)

摘要 (379) PDF (426KB) (58)

相关文章 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200506#RelatedCitationTab) | 多维度评价 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200506#MetricsTab)

重大突发疫情事件中的谣言识别 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200489)



(https://zrb.bjb.scut.edu.cn/fileup/1000-565X/PIC/2020-0489_pic.jpg)

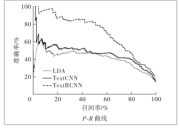
刘勘 黄哲英

2021, 49(1): 18-28. doi:10.12141/j.issn.1000-565X.200489 (https://doi.org/10.12141/j.issn.1000-565X.200489)

摘要 (485) PDF (1872KB) (92)

相关文章 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200489#RelatedCitationTab) | 多维度评价 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200489#MetricsTab)

基于半监督学习的涉及未成年人案件文书识别方法 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200513)



(https://zrb.bjb.scut.edu.cn/fileup/1000-565X/PIC/2020-0513_pic.jpg)

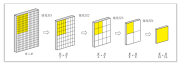
杨圣豪, 吴玥悦, 毛佳昕, 等

2021, 49(1): 29-38,46. doi:10.12141/j.issn.1000-565X.200513 (https://doi.org/10.12141/j.issn.1000-565X.200513)

摘要 (381) PDF (1190KB) (96)

相关文章 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200513#RelatedCitationTab) | 多维度评价 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200513#MetricsTab)

基于点云俯视图重映射的 3D 目标检测 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200373)


(https://zrb.bjb.scut.edu.cn/fileup/1000-565X/PIC/2020-0373_pic.jpg)

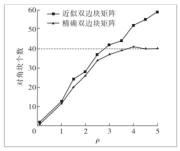
吴秋霞 黎玲敏

2021, 49(1): 39-46. doi:10.12141/j.issn.1000-565X.200373 (https://doi.org/10.12141/j.issn.1000-565X.200373)

摘要 (528) PDF (1263KB) (141)

相关文章 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200373#RelatedCitationTab) | 多维度评价 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200373#MetricsTab)

面向非随机缺失数据的协同过滤评分方法 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200210)


(https://zrb.bjb.scut.edu.cn/fileup/1000-565X/PIC/2020-0210_pic.jpg)

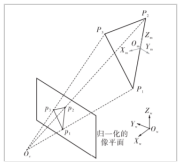
古万荣, 谢贤芬, 张子烨, 等

2021, 49(1): 47-57. doi:10.12141/j.issn.1000-565X.200210 (https://doi.org/10.12141/j.issn.1000-565X.200210)

摘要 (278) PDF (786KB) (64)

相关文章 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200210#RelatedCitationTab) | 多维度评价 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200210#MetricsTab)

透视三点问题的一种快速且稳定的代数解法 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200399)


(https://zrb.bjb.scut.edu.cn/fileup/1000-565X/PIC/2020-0399_pic.jpg)

耿庆华 刘伟铭

2021, 49(1): 58-64,73. doi:10.12141/j.issn.1000-565X.200399 (https://doi.org/10.12141/j.issn.1000-565X.200399)

摘要 (596) PDF (867KB) (81)

相关文章 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200399#RelatedCitationTab) | 多维度评价 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200399#MetricsTab)

通知公告

(https://zrb.bjb.scut.edu.cn/CN/news
2024-01-07

本刊2023年第10期被EI收录论文目录
(https://zrb.bjb.scut.edu.cn/CN/news
2023-12-28

本刊2023年第9期被EI收录论文目录
(https://zrb.bjb.scut.edu.cn/CN/news
2023-12-01

本刊2023年第8期被EI收录论文目录
(https://zrb.bjb.scut.edu.cn/CN/news
2023-10-31

更多... (../news/folde

精彩推荐

车路协同系统关键技术研究进展
(../10.12141/j.issn.1000-565X.2302

考虑道路负载均衡的码头多AGV无冲
(../10.12141/j.issn.1000-565X.2302

TSN中基于链路负载均衡的AVB流量
(../10.12141/j.issn.1000-565X.2207

一种面向中文拼写纠错的自监督预训
(../10.12141/j.issn.1000-565X.2300

多类型公交线路的运行可靠性优化策
(../10.12141/j.issn.1000-565X.2204

面向边缘特征的实时模板匹配方法
(../10.12141/j.issn.1000-565X.2207

更多>> (../article/sho

阅读排行

第49卷第1期
2021年1月

华南理工大学学报(自然科学版)
Journal of South China University of Technology
(Natural Science Edition)

Vol. 49 No. 1
January 2021

doi:10.12141/j.issn.1000-565X.200210

面向非随机缺失数据的协同过滤评分方法

古万荣[1] 谢贤芬[2] 张子烨[3] 毛宜军[1†] 梁早清[1] 何亦琛[1]

(1. 华南农业大学 数学与信息学院,广东 广州 510642;2. 暨南大学 经济学院,广东 广州 510632;
3. 华南理工大学 数学学院,广东 广州 510640)

摘 要:大多数评分预测研究都是基于缺失值是随机的假设。然而,实际的线上推荐系统的评分矩阵的缺失数据都是非随机的。对缺失数据的错误假设会导致有偏差的参数估计和预测。为了提高非随机缺失评分矩阵填补的准确度,文中深入分析了用户和物品的评分矩阵的内在原理,提出了通过行或列变换将用户和物品的评分矩阵转变为等价的双边块对角矩阵,再在不同的分区块中分别应用矩阵分解方法进行分解和评分预测的方法,使得局部数据更新和分解成为现实。在公测数据集上的实验结果显示,文中方法可以提高评分填补效果,有效地解决非随机评分缺失问题,从而提高推荐系统的预测准确率。变换后的分块矩阵在分布式处理实验中也获得了较好的加速比,说明文中方法具有较好的应用可扩展性。

关键词:矩阵分解;推荐系统;奇异值分解;评分预测

中图分类号:TP31 **文章编号**:1000-565X(2021)01-0047-11

随着互联网电子商务的发展,推荐系统显得日益重要。直接使用推荐结果,用户可以减少浏览时间、提升网购的满意度,商家可以提高转化率和利润率。自 Netflix 奖金设立以来,推荐技术的研究尤其在协同过滤算法方面的研究日益被学术界和工业界所关注[1]。

目前大多数推荐技术的研究都基于用户历史行为,或融合一些上下文因素,如时间周期、点击间隔等[2-3],比较少对用户和物品的协同矩阵的内在结构进行深入理解和剖析。因此,大多数推荐算法都是基于模型或基于近邻方法,利用评分预测矩阵等代数工具进行挖掘。这些方法大多基于这种假设:评分预测矩阵的缺失或用户物品关联的缺失是随机的。这种假设使得各种评分预测方法倾向于将已有数据和预测数据看作是同一分布的,这并不符合电子商务或社交网络中的兴趣组和社群分块的实际情况。因此,目前对推荐系统的研究仍然存在至少3个方面的问题:①关联数据的稀疏性[4]。这是在大数据挖掘中普遍存在的一个特点,在推荐系统研究中,这个问题也始终没有得到很好的解决。稀

收稿日期:2020-05-06

基金项目:国家重点研发计划项目(2017YFC1601701);广东省科技计划项目(2018A070712021);国家统计科学研究重点项目(2019LZ37);广东省哲学社会科学规划项目(GD18CXW01,GD19CGL34)

Foundation items: Supported by the National Key R&D Program of China (2017YFC1601701), the Science and Technology Planning Project of Guangdong Province (2018A070712021), the National Key Project of Statistical Science Research (2019LZ37) and the Philosophy and Social Sciences Planning Project of Guangdong Province (GD18CXW01, GD19CGL34)

作者简介:古万荣(1982-),男,博士,讲师,硕士生导师,主要从事互联网大数据处理与分析和推荐模型研究。E-mail:guwanrong@scau.edu.cn

† **通信作者**:毛宜军(1979-),男,博士,讲师,硕士生导师,主要从事大数据分析和生物信息学研究。E-mail:yijunmao@163.com

疏性使得轻微的变动或计算误差将导致结果的剧烈变动。②关联矩阵规模庞大导致算法可扩展性差[5]。在推荐系统中，许多算法依赖于关联矩阵的更新，每一次更新都要将关联矩阵重新计算或分解，由于关联矩阵较大，计算的复杂性导致这些研究成果很难在应用系统中使用。③算法缺乏对不同用户社区群的区分[6]。这就导致了有些具备特殊需求的用户长期得不到正确的兴趣挖掘，影响了整体推荐精度。

基于内容的推荐的基本原理在于挖掘用户的历史行为所关联的物品内容，即与内容是相关的[7]。基于协同过滤的推荐方法不需要分析内容，即通过用户和物品的协同关系来确定给定用户可能感兴趣的物品列表[8-9]。由于纯基于内容的推荐方法实际上是将推荐预测的关联性看作是非随机的，因此，这就容易解释为什么很多应用使用基于内容的方法比单纯使用协同过滤方法的准确度要高。协同过滤方法是推荐算法的另外一类算法，主要原理是通过用户和物品的关联协同性进行基于经验或基于模型的推荐预测。矩阵分解方法在推荐技术的研究中因具有很高的推荐准确度而受到学术界的广泛关注[10]。经典的方法有奇异值分解（SVD）法[11]和NMF非负矩阵分解（NMF）法[12-13]。在理论研究上，矩阵分解方法可以有效地挖掘用户和物品之间的潜在关联因素，预测的精度较高。在推荐系统实践中，每次用户新购买或点击时，将会对原始的用户和物品的二分图产生影响，即原始关联矩阵将会改变。在线系统中，用户购买物品的关联动作以及用户评价物品的评分矩阵是动态变化的。随着每一个数据的细微改变，都需要对相应的矩阵进行分解，导致该类算法很难在实际系统中得到应用[14]。此外，由于缺失数据的非随机性，出现了不同兴趣聚集和社团聚集现象。为此，如果能只对矩阵的局部变化做出局部分解，既能减轻计算负担，又能有效挖掘非随机缺失的评分数据，从而使该类算法在较准确的预测率情况下应用到实际的大型推荐系统中[15-17]。在应用系统开发中，也有人提出了间隔性增量式的分解，即不对用户的每一个动作做出即时的分解应答。这种做法虽然缓解了运算负担，但从本质上来说是以推荐模型的时效性作为代价的，而且仍然没有解决全局矩阵分解的大量运算问题。

在算法优化方面，文献［18-19］提出了使用矩阵聚类的方法来提高算法的效率，以提高应用的可扩展性。这类方法的基本原理是先对用户或物品的行列向量进行聚类，然后以聚类为单位进行计算，从而缩小了运算的规模和粒度。也有算法[20]提出了同时对行列向量进行聚类，这样进一步缩减了后续计算的规模，具有一定的效果。这类方法在应用中具有较强的可扩展性，但规定了某用户或某物品只能属于某一个类别，这跟实际情况并不完全符合。因为在现实的电子商务环境中，用户或物品往往从属于多个类，即便是从属于同一个类，参与度也不能对等看待。

在社交网络的研究中，基于图的社区发现可以等价表示为关联矩阵的近似双边块对角结构[21-22]。基于此思想，本文基于评分缺失非随机性假设，提出了先将用户和物品的评分矩阵转化为近似双边块对角矩阵，再在近似双边块对角矩阵上运用矩阵分解方法进行评分预测的方法，并运用协同过滤算法验证其有效性。

1 基于矩阵分解的推荐技术

推荐技术的研究往往集中于用户和物品之间的协同关系，最后预测用户和物品之间的兴趣关系。一般有Top@n推荐和评分预测两种应用问题。一般而言，Top@n推荐的研究需要建立用户和物品之间的兴趣度函数，以计算特定用户对特定物品之间的兴趣度或购买意愿。评分预测问题也可称为缺失评分填补问题或矩阵填补问题，常使用代数上已证明的矩阵分解方法来获得填补值。

矩阵分解一般是指将一个矩阵，即推荐系统中的评分预测矩阵，分解为两个或多个矩阵的乘积。在物理意义上可以理解为，用户对物品的喜好，都是基于对物品的特殊因子的喜好，不同的因子有不同的权重，对于新物品，如果它具有用户喜好的因子，则该物品被推荐的概率更大。矩阵分解式如下：

$$\boldsymbol{R}_{m\times n}\approx\boldsymbol{P}_{m\times k}\times\boldsymbol{Q}_{k\times n}=\hat{\boldsymbol{R}}_{m\times n} \tag{1}$$

那么问题就转化为如何求得$\boldsymbol{P}_{m\times k}$和$\boldsymbol{Q}_{k\times n}$矩阵中的元素问题，该问题可以通过转化为回归问题来解决。对于$\boldsymbol{R}_{m\times n}$矩阵中的每个元素，可以表示为如下形式：

$$e_{i,j}^2=(r_{i,j}-\hat{r}_{i,j})^2=(r_{i,j}-\sum_{k=1}^{K}p_{i,k}q_{k,j})^2 \tag{2}$$

式中，$r_{i,j}$和$\hat{r}_{i,j}$分别为矩阵$\boldsymbol{R}_{m\times n}$元素的实际评分值和通过算法预估的评分值。基于式（2）的平方损失函数，可以通过梯度下降法求最优解，许多经典的矩阵分解方法都是基于该原理，即通过求解最小化的平方损失函数，使用梯度下降法迭代获得矩阵

分解结果，如 SVD[23]、NMF[24]等。通过矩阵分解求解后，可以将原始矩阵的缺失值求出来，从而实现基于评分预测的推荐应用。

2　矩阵分块变换

2.1　基本定义

为了方便说明问题，本文将给出双边块矩阵的相关通用定义。

定义 1（双边块对角矩阵）　矩阵 $\boldsymbol{X}$ 被称为双边块对角矩阵，当且仅当该矩阵可以通过变换变成如下形式：

$$\boldsymbol{X}=\begin{bmatrix}\boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \cdots & \boldsymbol{A}_{1k} & \boldsymbol{A}_{1B}\\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \cdots & \boldsymbol{A}_{2k} & \boldsymbol{A}_{2B}\\ \vdots & \vdots & & \vdots & \vdots\\ \boldsymbol{A}_{k1} & \boldsymbol{A}_{k2} & \cdots & \boldsymbol{A}_{kk} & \boldsymbol{A}_{kB}\\ \boldsymbol{A}_{B1} & \boldsymbol{A}_{B2} & \cdots & \boldsymbol{A}_{Bk} & \boldsymbol{A}_{BB}\end{bmatrix}=\begin{bmatrix}\boldsymbol{D}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{C}_1\\ \boldsymbol{0} & \boldsymbol{D}_2 & \cdots & \boldsymbol{0} & \boldsymbol{C}_2\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{D}_k & \boldsymbol{C}_k\\ \boldsymbol{R}_1 & \boldsymbol{R}_2 & \cdots & \boldsymbol{R}_k & \boldsymbol{B}\end{bmatrix}\tag{3}$$

式中：每个子矩阵 $\boldsymbol{D}_i$ 被称为“对角块”；子矩阵 $\boldsymbol{R}=[\boldsymbol{R}_1,\boldsymbol{R}_2,\cdots,\boldsymbol{R}_k,\boldsymbol{B}]$ 和 $\boldsymbol{C}=[\boldsymbol{C}_1^{\mathrm{T}},\boldsymbol{C}_2^{\mathrm{T}},\cdots,\boldsymbol{C}_k^{\mathrm{T}},\boldsymbol{B}^{\mathrm{T}}]$ 称为该矩阵的“双边”。由于这两条“双边”并非向量，而是一个子矩阵，因此又称为“双边块”。子矩阵 $\boldsymbol{D}_i$ 通过同样算法也可以被排列为同样的双边块。从该公式易得，当 $\boldsymbol{R}$ 和 $\boldsymbol{C}$ 都为零矩阵时，该矩阵 $\boldsymbol{X}$ 则为块对角矩阵。

定义 2（近似双边块对角矩阵，ABBDF）　当矩阵 $\boldsymbol{X}$ 只存在若干个非 0 值时，称矩阵 $\boldsymbol{X}$ 是一个近似双边块对角矩阵，通过行或列变换可以变成如下形式：

$$\boldsymbol{X}=\begin{bmatrix}\boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \cdots & \boldsymbol{A}_{1k} & \boldsymbol{A}_{1B}\\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \cdots & \boldsymbol{A}_{2k} & \boldsymbol{A}_{2B}\\ \vdots & \vdots & & \vdots & \vdots\\ \boldsymbol{A}_{k1} & \boldsymbol{A}_{k2} & \cdots & \boldsymbol{A}_{kk} & \boldsymbol{A}_{kB}\\ \boldsymbol{A}_{B1} & \boldsymbol{A}_{B2} & \cdots & \boldsymbol{A}_{Bk} & \boldsymbol{A}_{BB}\end{bmatrix}=\begin{bmatrix}\boldsymbol{D}_1 & \boldsymbol{A}_{12} & \cdots & \boldsymbol{A}_{1k} & \boldsymbol{C}_1\\ \boldsymbol{A}_{21} & \boldsymbol{D}_2 & \cdots & \boldsymbol{A}_{2k} & \boldsymbol{C}_2\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ \boldsymbol{A}_{k1} & \boldsymbol{A}_{k2} & \cdots & \boldsymbol{D}_k & \boldsymbol{C}_k\\ \boldsymbol{R}_2 & \boldsymbol{R}_2 & \cdots & \boldsymbol{R}_k & \boldsymbol{B}\end{bmatrix}\tag{4}$$

与定义 1 相比，定义 2 多了一个“近似”的表述，即近似双边块对角矩阵允许包含一些非 0 值的散点。

定义 3（矩阵的密度）　设 $\boldsymbol{X}$ 是一个 $m\times n$ 的矩阵，即 $\boldsymbol{X}$ 的矩阵元素个数为 $|m\times n|$，设 $f(\boldsymbol{X})$ 表示 $\boldsymbol{X}$ 中非 0 元素的个数，则矩阵 $\boldsymbol{X}$ 的密度 $\rho(\boldsymbol{X})=\dfrac{f(\boldsymbol{X})}{|m\times n|}$。

2.2　分块矩阵变换算法

在社区发现的研究中，高密度聚集的社区意味着兴趣爱好相似。高密度的社区可以为下一步基于矩阵分解的协同过滤评分预测提供数据基础。为清晰起见，本文提出了精确双边块对角化转换和近似双边块对角化转换算法。

（1）精确矩阵分块算法

在双边块对角化的变换过程中，基本操作就是迭代执行行和列的交换，使之形成双边块对角化形式。其中一个重要的步骤就是将某些行或列向量排列到双边。

该过程等价于在用户和物品的关联二分图中基于点割集的分割。一般而言，一个图的点割集可能有很多个，因此最优化的方式是找到最小的点分割集。在图论研究中，已知找最小的点割集是 NP 难问题。由于该问题的求解具有实际应用意义，因此人们提出了许多基于启发式的方法来实现图的点分割，如基于层次分析、谱分析和基于核函数的方法等。

本文算法在获得矩阵对角化时，将使用到子矩阵密度，单层双边块对角矩阵和对角块的平均密度计算方法为 GetDensity $(\boldsymbol{X},G)$。该算法描述如下：

GetDensity$(\boldsymbol{X},G)$ //获取密度

{输入：评分矩阵 $\boldsymbol{X}$ 及其二分图 $G(V,\varepsilon)=(R\cup C,\varepsilon)$，$V=R\cup C$，$R$ 和 C 分别是行和列的集合。

输出：平均密度 $\bar{\rho}$。

$\Gamma_v\leftarrow\{V_1,V_2,\cdots,V_k;V_s\}\leftarrow\mathrm{GP}(G)$;

//使用点割集函数 GP( )获得分割集

将矩阵 $\boldsymbol{X}$ 的行按 $R_1R_2\cdots R_kR_s$ 的顺序排列;

将矩阵 $\boldsymbol{X}$ 的列按 $C_1C_2\cdots C_kC_s$ 的顺序排列;

//根据行列排序后，$\boldsymbol{D}_i$ 互不相交

return $\bar{\rho}(\boldsymbol{D}_1,\boldsymbol{D}_2,\cdots,\boldsymbol{D}_k)$;

//$\boldsymbol{D}_i$ 对应的顶点集是 $V_i=R_i\cup C_i$}

在该算法的迭代过程中，应尽量让排列后的对角块的平均密度高于原始矩阵密度。

双边块对角化变换算法描述如下:

TwoSideBlock($\boldsymbol{X}$,G,ρ) //双边块对角化变换

{输入:目标密度ρ、评分矩阵$\boldsymbol{X}$及其二分图$G(V,\varepsilon)=(R\cup C,\varepsilon)$,$V=R\cup C$,$R$和$C$分别是行和列的集合。

输出:排列好的双边块对角矩阵$\boldsymbol{X}$。

$\rho_X\leftarrow\rho(\boldsymbol{X})$; //将矩阵平均密度用$\rho_X$保存

if $\rho_X<\rho$ then

//当平均密度小于预设目标密度ρ时

$\bar{\rho}\leftarrow$GetDensity($\boldsymbol{X}$,G);

//获取当前密度用$\bar{\rho}$保存块平均密度

if $\bar{\rho}>\rho_X$ then //当$\bar{\rho}<=\rho_X$时,递归调用停止

for 每一个对角块$\boldsymbol{D}_i$ do

TwoSideBlock($\boldsymbol{D}_i$,G_{V_i},ρ);

//对每个对角块$\boldsymbol{D}_i$递归调用变换函数

end

end

end}

在该算法中,目标密度ρ需要在函数调用时设定。当算法迭代计算达到这个阈值时终止。该阈值如果设置过高,则可能会导致迭代计算不会停止,针对这一问题,可以通过对迭代次数进行控制:在代码中针对迭代次数设置一个计时器,当迭代次数超过给定阈值时,不管密度是否达到目标值,都停止迭代。

(2) 近似矩阵分块算法

与精确双边块对角化算法不同,在近似对角块变换方法中,本文通过抽取向量到边上来进一步提高密度,直到密度达到要求或迭代次数超过预设值。算法描述如下:

ApprTwoSideBlock($\boldsymbol{X}$,G,ρ)

//近似双边块对角化变换

{输入:目标密度ρ、评分矩阵$\boldsymbol{X}$及其二分图$G(V,\varepsilon)$,$V=R\cup C$,R、C分别是行和列的集合。

输出:排列好的近似双边块对角矩阵$\boldsymbol{X}$。

if $\rho_X>\rho$ then

//当平均密度大于目标密度时,算法终止

return;

else

$\Gamma_e\leftarrow\{V_1,V_2,\cdots,V_k\}\leftarrow$GP($G$);

//使用点割集函数获得分割集

将矩阵$\boldsymbol{X}$的行按$R_1R_2\cdots R_k$的顺序排列;

将矩阵$\boldsymbol{X}$的列按$C_1C_2\cdots C_k$的顺序排列;

//根据行列排序后,$\boldsymbol{D}_i$互不相交

$\{V'_1,V'_2,\cdots,V'_k;V'_s\}\leftarrow$ImprGetDensity($\boldsymbol{X}$,$G$,$\Gamma_e$);

//使用改进的密度函数生成分块集

for 每一个对角块$\boldsymbol{D}_i$ do

ApprTwoSideBlock($\boldsymbol{D}_i$,G_{V_i},ρ);

//对每个对角块$\boldsymbol{D}_i$递归调用近似分块算法

end

end}

该算法调用了ImprGetDensity($\boldsymbol{X}$,G,Γ_e)函数,该函数是改进的密度计算方法。该函数从对角块中选一处最大限度提高平均密度的向量并移动到边。最简单的做法是逐个穷举移动,并计算平均密度值,选取最大限度提高密度的移动方法。但在实际应用中,往往是专注于最小密度向量,即对子对角块中具有较少非0的向量进行缓存。改进的密度计算算法描述如下:

ImprGetDensity($\boldsymbol{X}$,G,Γ_e)

{输入:评分矩阵$\boldsymbol{X}$及其二分图$G(V,\varepsilon)$,以及图$G(V,\varepsilon)$中的GP(G)函数的运算结果Γ_e,$V=R\cup C$,R和C分别是行和列的集合。

输出:更好地提升密度的图分割结果Γ'。

$\{V'_1,V'_2,\cdots,V'_k;V'_s\}\leftarrow\{V_1,V_2,\cdots,V_k;\theta\}$;

//将目前的分块用临时变量存储

while $\rho(\boldsymbol{D}_1,\boldsymbol{D}_2,\cdots,\boldsymbol{D}_k)<\rho(\boldsymbol{X})$ do

//当目前的分块平均密度小于整个矩阵密度时

$x'\leftarrow0$,$i'\leftarrow0$,$\bar{\rho}\leftarrow0$; //初始化临时变量

for 每一个对角块$\boldsymbol{D}_i$ do

for $\boldsymbol{D}_i$的每一个行和列向量 do

//对每个对角块$\boldsymbol{D}_i$及其行列向量进行迭代

$$\bar{\rho}\leftarrow\frac{\sum_{j=1}^{k}n(\boldsymbol{D}_j)-n(x(\boldsymbol{D}_i))}{\sum_{j=1}^{k}\text{area}(\boldsymbol{D}_j)-\text{area}(x(\boldsymbol{D}_i))};$$

//选择对角块移除后的块平均密度,$n(\boldsymbol{D})$为$\boldsymbol{D}$的非0个数,$x(\boldsymbol{D})$为$\boldsymbol{D}$中对角块的非0个数,area($\boldsymbol{D}$)为$\boldsymbol{D}$的区域个数。

if $\bar{\rho}>\bar{\rho}'$ do//最大限度提升块平均密度

$\boldsymbol{x}'\leftarrow\boldsymbol{x}$,$i'\leftarrow i$,$\bar{\rho}'\leftarrow\bar{\rho}$;

//记录相关信息

end

end

end

将 $\boldsymbol{x}'$向量交换到双边；

$V'_{i'} \leftarrow V'_{i'} - \{\text{node}(\boldsymbol{x}')\}$；//从 $V'_{i'}$中移除相关节点；node $(\boldsymbol{x}')$为 $V'_{i'}$中对应向量的第 i' 个节点

$V'_s \leftarrow V'_s \cup \{\text{node}(\boldsymbol{x}')\}$；

//将相关的节点加入到 V'_s中

end

return $\Gamma' \leftarrow \{V'_1, V'_2, \cdots, V'_k; V'_s\}$；

//记录分块信息并返回。}

2.3　基于分块矩阵的协同过滤

矩阵分解算法的执行并没有限定矩阵形态，因此将用户和物品的关联矩阵或评分矩阵进行转换后，仍然可以使用基于矩阵分解的协同过滤评分预测方法。也就是说，矩阵变换前后，其物理特性并没有改变。但在矩阵分解挖掘过程中，对具有高度兴趣社群聚集的分块子矩阵进行分解挖掘，可以更聚焦于特定兴趣的用户群，减少全局无关因素干扰，同时也提高了评分预测的整体处理效率。假设只有单核处理器对数据进行处理，矩阵分块后规模变小，子矩阵分解时间呈指数级下降，虽然需要依次处理多个子矩阵，但多个子矩阵的轮换处理仅仅增加线性倍数时间。综合考虑，本文采用先分块再进行分解的方法仍然具有很大的效率优势。双边块对角矩阵的基本结构如图 1 所示。

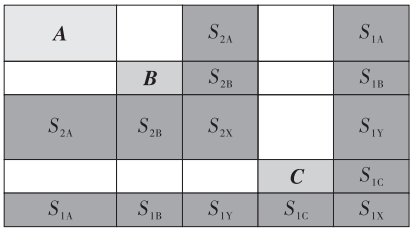


图 1　双边块对角矩阵的基本结构

Fig. 1　Basic structure of two-sided block diagonal matrix

如图 1 所示，对于每一个对角块，可以通过双边拼接方法来重建子用户物品群，即可以通过挑选部分用户和物品及其关联集合来进行协同过滤推荐预测。例如可以通过 $\boldsymbol{A}$、$\boldsymbol{B}$、$\boldsymbol{C}$对角块重构3个子矩阵，并忽略在非对角块上的零星非 0 值点，结果如图 2 所示。

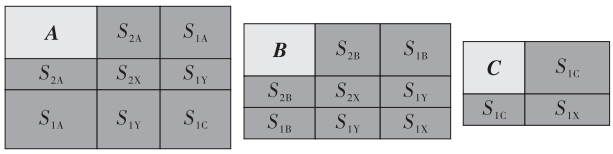


图 2　根据双边块对角型矩阵抽取的子矩阵

Fig. 2　Submatrices extracted from two-sided block diagonal matrix

令 $\boldsymbol{X} \in \mathbf{R}^{m \times n}$为一个稀疏矩阵，$\boldsymbol{U} \in \mathbf{R}^{m \times r}$，$\boldsymbol{V} \in \mathbf{R}^{r \times n}$，则矩阵分解算法可以形式化表示为 $P = (f, D_{\boldsymbol{w}}, C, R)$，其中，预测函数 f 是自变量和因变量都是 $m \times n$ 的矩阵：$\mathbf{R}^{m \times n} \rightarrow \mathbf{R}^{m \times n}$；数据权重矩阵$\boldsymbol{w} \in \mathbf{R}^{m \times n}$；$D_{\boldsymbol{w}}(\boldsymbol{X}, f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}})) \geqslant 0$ 为损失函数，用于记录预测值$f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}})$和原始值 $\boldsymbol{X}$ 之间的误差；C 为对分解后的矩阵 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 的约束；$R(\boldsymbol{U}, \boldsymbol{V}) \geqslant 0$ 是正则化项。因此，在矩阵分解模型中，需要求解以下优化函数：

$$\arg\min_{(\boldsymbol{U}, \boldsymbol{V} \in C)} \left[D_{\boldsymbol{w}}(\boldsymbol{X}, f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}})) + R(\boldsymbol{U}, \boldsymbol{V}) \right],$$

其中，函数 $D(\cdot, \cdot)$返回值是评分预测矩阵中各个元素的评估损失的累加和。

3　实验与结果分析

3.1　实验数据集

在本文实验中，使用了公测数据集 MoiveLens-100K（ML100K）和 MovieLens-1M（ML1M）、雅虎音乐数据集（YM）以及大众点评数据集（DP），如表 1 所示。这些数据集都遵循着观看或购买后才可以点评的较高评分门槛，因此具备一定的兴趣社群特性。例如，一个用户观看电影有一定的偏向性，因此他缺失的评分就具有非随机的特点。

3.2　参数调整及分析实验

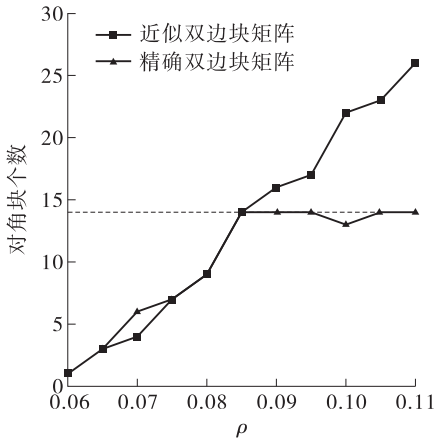
不同的目标密度会对生成的分块矩阵有不同的分布结果，对协同过滤评分预测精度也具有重要的意义。如果目标密度设置过低，则有些用户和物品

表 1　实验使用的数据集

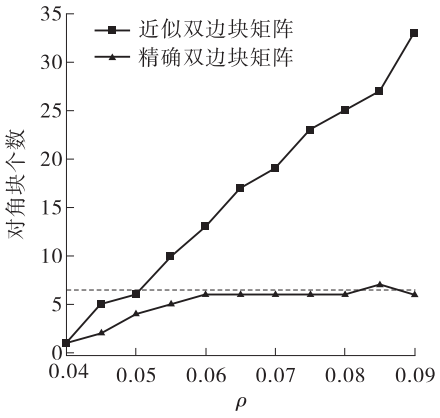
Table 1　Data sets used in experiment

数据集	用户	物品	评分	每用户评分数	每物品评分数	密度
ML100K	934	1682	100000	106	59	0.06300
ML1M	6040	3952	1000209	166	253	0.04190
DP	11857	22365	510551	43	23	0.00193
YM	1000990	624961	262810175	263	421	0.00042

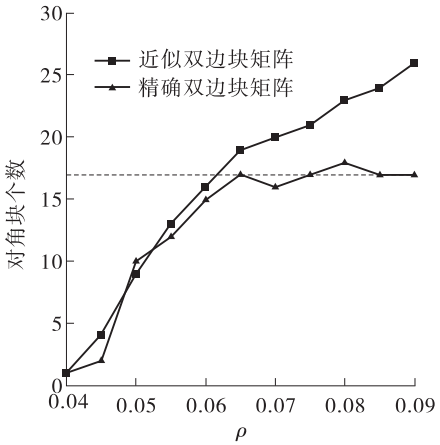
的关联社区群不能充分挖掘出来；如果目标密度设置过高，则会出现较多零散社区群，也同样不利于协同过滤预测。目标密度 ρ 与对角块个数的关系如图3所示。从图中可知：随着 ρ 的增大，精确双边块矩阵的对角块个数会增加，然后进入平稳状态，其直观的原因是，当目标密度达到一定的上限后将无法再上升，即无法再进行变换；近似双边块矩阵的对角块个数基本上随着 ρ 的增大而增长，说明通过将低密度的行或列移动到双边来可进一步提升密度。



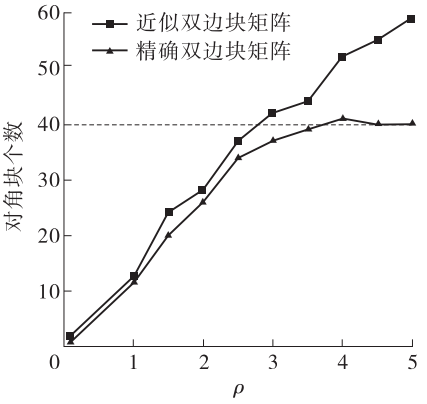
(a) ML100K



(b) ML1M



(c) DP



(d) YM

图3 不同数据集的目标密度与对角块个数的分布统计

Fig. 3 Distribution statistics of target density and number of diagonal blocks for different data sets

3.3 实验结果分析

为了对提出的基于对角化双边块矩阵协同过滤算法的预测精度进行评价，本文使用6种基线方法（SVD[23]、NMF[24]、PMF[25]、MMMF[26]、Faster-SVD[27]、ISVD[28]）作为参考方法进行对比。SVD是一种经典的方法，本文实验中使用Frobenius范数和交替最小化方差来进行训练和优化。NMF与SVD类似，也使用Frobenius范数。PMF是一种常用的概率分解模型，在本文实验中采用基于马尔可夫的贝叶斯模型。MMMF是一种经典的基于最大间隔的矩阵分解模型。Faster-SVD方法基于奇异值阈值分割（SVT）算法并结合子空间回收技术进行加速，该方法不仅在预测速度上有了较大的提升，而且还融合了处理稀疏矩阵的两种快速随机化算法（RSVD-PI和RSVD-DBKI），使得该方法在评分预测推荐方面有较强的优势。ISVD方法基于SVD算法原理，但在数据输入时整合了用户和物品邻域进行优化处理，对稀疏矩阵具有较强的准确性和稳定性。这些基线方法中，Faster-SVD和ISVD代码来自于其作者，只提供了单机版本。本文在此基础上对其进行了分布式并行化的一些修改调整，以适应本文实验。在评分预测中，最常用的评测指标是根均方差（RMSE），令 r_i 和 $\hat{r}_i$ 分别为真实评分和估计评分，则

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i-\hat{r}_i)^2}\ 。$$

本文实验采用了10折交叉验证，即将实验数据切分为10份，其中1份作为测试集，另外9份作为训练集，然后轮换其他另外1份作为测试集，剩余的作为训练集，轮询重复10次后，获得的准

确度平均值作为实验最终结果。

在矩阵分解算法中，分解因子的个数是预先设定的。如果该个数太小，则不足以近似表示原始关联矩阵的有效信息；如果该个数太大，一方面模型复杂耗时，另一方面容易产生过拟合问题。

首先使用 ML1M 数据集，$\bar{\rho}=0.055$。这样，原始矩阵就变换为具有 4 个对角块的多层双边块矩阵，如表 2 所示。

表 2　ML1M 的对角块数据

Table 2　Diagonal block data of ML1M

对角块	用户数	物品数	评分	密度
$\bar{X}_1$	1517	2591	118679	0.0302
$\bar{X}_2$	1693	3208	255896	0.0471
$\bar{X}_3$	1543	3476	464783	0.0867
$\bar{X}_4$	1350	3294	201147	0.0452

本文设置步长为 5，从 5 到 100 调整矩阵分解的因子个数 r。每个矩阵分解算法有以下两组实验：①在原始的关联矩阵 $\boldsymbol{X}$ 上进行矩阵分解，分解因子个数设置为 r，以此获得相应的评分预测准确度 RMSE 值。②在近似局部化矩阵分解框架下对 4 个对角块进行矩阵分解，分解因子个数设置为 r，以此获得相应的评分预测准确度 RMSE 值。在矩阵 $\boldsymbol{X}$ 上采用交叉验证方法确定参数，SVD 和 NMF 算法的正则化系数 $\lambda=0.065$，PMF 的系数 $\lambda_U=\lambda_V=0.002$，MMMF 的 $C=1.5$。各种方法的评分预测精度和分解因子个数的关系如图 4 所示，其中 LMF 表示使用本文提出的近似双边块对角矩阵的分布式矩阵分解方法。从图中可知：本文方法能够提高基于矩阵分解的协同过滤评分预测的精度；分解因子个数越少，预测精度提升越明显，这说明在本文的局部化双边块矩阵分解框架下的小规模矩阵分解可以达到较好的效果。

目标密度取不同值时对预测精度的影响如图 5 所示。从图中可知，理想的目标密度参数随着数据集的不同而不同，每条虚线以下的曲线部分表示 RMSE 较小，意味着评分预测的准确度较高，反之，在虚线以上的曲线部分则表示 RMSE 较大，预测的准确度较低。因此，通过合适但并不苛刻的密度设置，本文方法可以提升评分预测的精度。但目标密度的设定需要遵循合适的经验规律，目标密度设置过高会导致预测精度降低。其直观的原因在于，本文方法在密度设置过高时会形成较多的零散碎块，从而影响最终的矩阵分解的评分预测准确度。

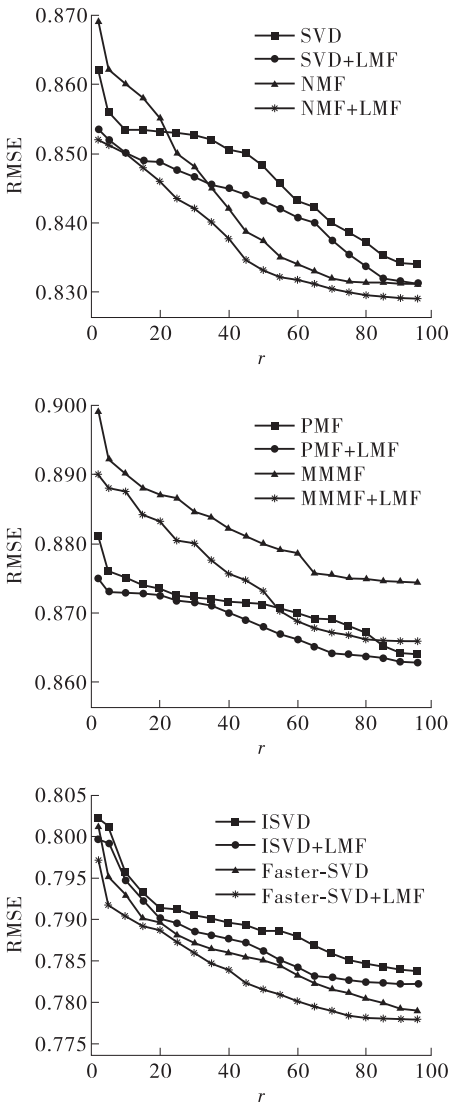


图 4　评分预测准确度和因子个数的关系

Fig. 4　Relationship between score prediction accuracy and number of factors

如表 3 所示，不同的目标密度参数设置会形成不同的矩阵变换数据分布，如形成不同的区块个数以及用户和物品分布。当目标密度设定过大（$\bar{\rho}\geqslant 0.1$）时，对角块的用户数只有几百个。然而，本文方法并不会因为经验设定而影响其可扩展性，实验结果表明，在目标密度非常大的取值范围内，本文方法对提升评分预测精度都具有明显的效果，如表 4 所示。从表 4 可知，在 ML 和 DP 数据集上，基于本文近似双边块对角矩阵方法的实验结果的统计标准差非常小（小于等于 0.002），在 YM 上的标准差小于等于 0.01。这说明本文方法具有很好的

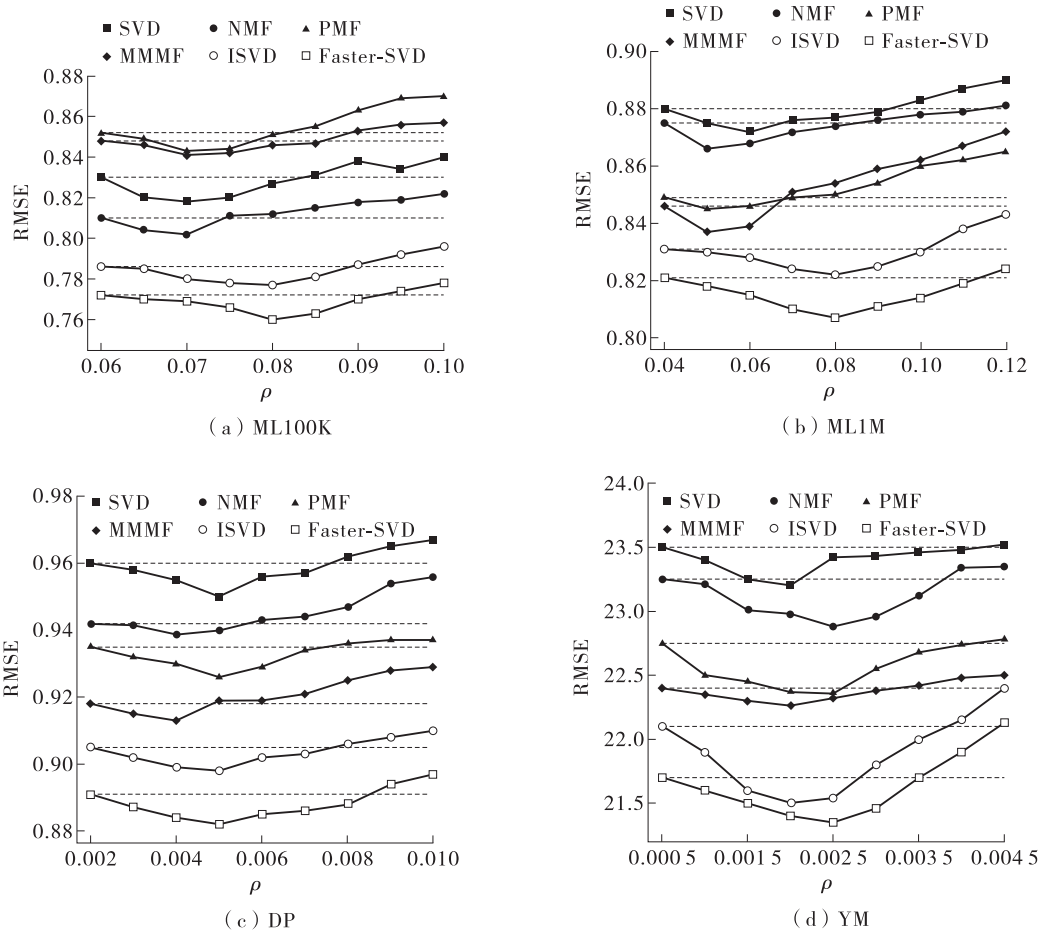


图5 目标密度和预测精度的关系

Fig. 5 Relations hip between target density and prediction precision

表3 ML1M在不同目标密度值下的对角块统计

Table 3 Diagonal block statistics of ML1M with different values of target density

$\bar{\rho}$	k	用户数均值	物品数均值
0.047	2	3121	3201
0.051	4	1623	3212
0.061	8	801	3166
0.070	16	398	3156
0.082	32	219	3123
0.110	64	130	3102
0.130	128	79	3120
0.159	256	59	3109

稳定性。其直观的原因是，一方面在原始矩阵中采用近似双边块矩阵变换方法后的分解具有良好的兴趣聚集效果，其次是在双边块对角矩阵上的分解比直接在原始矩阵上分解具有更好的兴趣挖掘效果。基于改进SVD的ISVD和Faster-SVD在原始矩阵的分块处理的实验结果比传统的SVD等算法好，在使用本文先分块再分解的方法之后，仍然具有较优的预测精度。这表明，本文方法和特定的矩阵分解算法的独立性较好，可以较方便地装配在任何矩阵分解算法中，帮助其提高推荐和评分填补的效果。

本文方法先对用户和物品的评分矩阵进行近似双边块对角化变换，再对对角块进行单独的矩阵分解。这样做的好处在于：①可以局部进行矩阵分解，因此当用户单独对物品进行购买或评分后，推荐系统无需针对整个关联矩阵进行计算；②可以对每个对角矩阵进行分布式处理，从而提高推荐系统的并行化效率。本文方法的步骤如下：

1）将关联矩阵 $\boldsymbol{X}$ 转化为一个多层双边对角矩阵；

2）将原始矩阵切分为8个对角块，对每个对角块利用独立的处理核进行分解，使其互不干扰；

3）基于分解结构将其合并为近似预测原始关联矩阵。

在实验中记录每个步骤所消耗的时间，并统计

其分布式系统的加速比（对近似双边块对角矩阵的分布式分解与对原矩阵的直接分解的运行时间之比）。为了方便对比分析，每个矩阵分解因子和前面实验相同，即 $r=60$。实验结果如表 5 所示。

表 4　各种矩阵分解方法的参数和预测结果

Table 4　Parameters and prediction results of various matrix decomposition methods

方法	ML100K				ML1M				DP				YM			
	RMSE		$\bar{\rho}$	k	RMSE		$\bar{\rho}$	k	RMSE		$\bar{\rho}$	k	RMSE		$\bar{\rho}$	k
	原矩阵	ABBDF			原矩阵	ABBDF			原矩阵	ABBDF			原矩阵	ABBDF		
SVD	0.9250	0.9198	0.09	3	0.8598	0.8354	0.06	3	0.9365	0.9098	0.0037	3	23.0240	21.9860	0.0053	12
NMF	0.9249	0.9019	0.09	3	0.8555	0.8245	0.06	3	0.9421	0.9147	0.0045	4	24.2540	22.4780	0.0062	19
PMF	0.9603	0.9424	0.09	3	0.8874	0.8598	0.06	3	0.9754	0.9496	0.0045	4	23.6540	21.4980	0.0031	7
MMMF	0.9923	0.9635	0.09	3	0.8912	0.8546	0.07	9	0.9521	0.9257	0.0037	3	24.5120	22.8790	0.0041	10
ISVD	0.8546	0.8236	0.09	3	0.8874	0.8452	0.07	9	0.9230	0.9154	0.0035	4	22.9810	21.6320	0.0043	10
Faster-SVD	0.8423	0.8215	0.09	3	0.8756	0.8359	0.07	9	9.2102	0.9121	0.0041	4	22.6100	21.0150	0.0043	10

表 5　各种矩阵分解方法的实验结果

Table 5　Experimental results of various matrix decomposition methods

方法	ML100K			ML1M			DP			YM		
	运行时间/min		加速比	运行时间/min		加速比	运行时间/min		加速比	运行时间/h		加速比
	直接分解	分布式分解		直接分解	分布式分解		直接分解	分布式分解		直接分解	分布式分解	
SVD	24.1	8.2	2.93	198.2	39.8	4.98	152.6	34.8	4.39	6.33	1.18	5.36
NMF	9.6	4.2	2.29	90.8	19.5	4.66	70.2	15.3	4.59	4.96	0.97	5.11
PMF	52.5	9.9	3.78	245.8	59.4	4.14	201.3	43.8	4.60	8.03	1.32	6.08
MMMF	20.1	5.1	5.30	2.0	30.2	3.97	50.3	9.7	5.19	40.20	6.15	6.53
ISVD	21.6	9.4	2.29	198.8	38.6	5.15	153.6	33.1	4.64	6.89	1.68	4.10
Faster-SVD	20.5	8.6	2.38	183.4	33.5	5.47	135.6	30.8	4.40	6.01	1.56	3.85

从表 5 可知，本文提出的近似双边块对角矩阵对分布式处理具有显著的加速效果，其原因在于双边块对角矩阵可以直接拆解，互不影响，这在推荐系统应用的海量数据处理中具有重要的现实意义。经典的矩阵分解算法（SVD、NMF、PMF 和 MMMF）比 ISVD 和 Faster-SVD 具有较高的加速比，其原因可能是经典算法的辅助处理步骤较少，可较大限度解决单纯使用基于分而治之思想进行分布式计算带来的重复计算问题。但不失一般性，本文提出的分块方法仍然可以很容易应用于分布式处理中。

4　结论与展望

为有效挖掘用户兴趣社群，降低矩阵分解或其他协同过滤算法的运算时间，本文基于缺失评分的非随机性特点，提出了先将用户和物品的评分矩阵变换为近似双边块对角矩阵，再对对角块进行单独的矩阵分解的方法。在公测数据集上的实验结果表明：基于双边块对角变换的矩阵分解方法具有更高的评分预测精度，应用在分布式处理中具有较高的加速比和良好的可扩展性。本文提出的方法的主要优点如下：

（1）本文方法先将原始矩阵变换为分块的矩阵，从而可以针对密度聚集的子矩阵进行深入分析，有效挖掘用户兴趣社群并缓解数据稀疏性的影响；

（2）本文方法的局部计算特性，使得实际系统的数据更新不需要对整个大矩阵进行重新分解，只需要在更新点所在的高密度块进行处理，从而提高了应用的可扩展性；

(3) 各高密度子矩阵的协同过滤和矩阵分解互不影响，使得算法的生成矩阵更容易并行化，提高了整体系统的预测效率。

未来将深入研究基于分块的矩阵分解的其他协同过滤算法，使得现有应用系统能无缝对接本文方法，从而实现其应用扩展。

参考文献：

[1] BEEL J, GIPP B, LANGER S, et al. Paper recommender systems: a literature survey [J]. International Journal on Digital Libraries, 2016, 17(4): 305-338.

[2] WEST J D, SMITH W I, BERGSTROM C T. A recommendation system based on hierarchical clustering of an article-level citation network [J]. IEEE Transactions on Big Data, 2016, 2(2): 113-123.

[3] LIU X, SONG Y, AGGARWAL C, et al. BiCycle: item recommendation with life cycles [C] // Proceedings of 2017 IEEE International Conference on Data Mining. New Orlens: IEEE, 2017: 297-306.

[4] NIKOLAKOPOULOS A N, KOUNELI M A, GAROFALAKIS J D. Hierarchical itemspace rank: exploiting hierarchy to alleviate sparsity in ranking-based recommendation [J]. Neurocomputing, 2015, 163: 126-136.

[5] YAO L, SHENG Q Z, NGU A H H, et al. Things of interest recommendation by leveraging heterogeneous relations in the Internet of things [J]. ACM Transactions on Internet Technology, 2016, 16(2): 1-25.

[6] WANG J, SUN J, LIN H, et al. Convolutional neural networks for expert recommendation in community question answering [J]. Science China Information Sciences, 2017, 60(11): 110102/1-9.

[7] LOPS P, JANNACH D, MUSTO C, et al. Trends in content-based recommendation [J]. User Modeling and User-Adapted Interaction, 2019, 29(2): 239-249.

[8] WEI J, HE J, CHEN K, et al. Collaborative filtering and deep learning based recommendation system for cold start items [J]. Expert Systems with Applications, 2017, 69: 29-39.

[9] YANG Z, WU B, ZHENG K, et al. A survey of collaborative filtering-based recommender systems for mobile Internet applications [J]. IEEE Access, 2016, 4: 3273-3287.

[10] ALMUTAIRI F M, SIDIROPOULOS N D, KARYPIS G. Context-aware recommendation-based learning analytics using tensor and coupled matrix factorization [J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(5): 729-741.

[11] ZHOU X, HE J, HUANG G, et al. SVD-based incremental approaches for recommender systems [J]. Journal of Computer and System Sciences, 2015, 81(4): 717-733.

[12] ALGHAMEDY F, ZHANG J. Enhance NMF-based recommendation systems with social information imputation [C] // Proceedings of the 3rd International Conference on Data Mining & Knowledge Management. Dubai: AIRCC, 2018: 37-54.

[13] ALGHAMEDY F, ZHANG J. Imputation strategies for cold-start users in NMF-based recommendation systems [C] // Proceedings of 2019 the 3rd International Conference on Information System and Data Mining. New York: ACM, 2019: 119-128.

[14] CUI L, HUANG W, YAN Q, et al. A novel context-aware recommendation algorithm with two-level SVD in social networks [J]. Future Generation Computer Systems, 2018, 86: 1459-1470.

[15] WEI S, ZHENG X, CHEN D, et al. A hybrid approach for movie recommendation via tags and ratings [J]. Electronic Commerce Research and Applications, 2016, 18: 83-94.

[16] CHU W T, TSAI Y L. A hybrid recommendation system considering visual information for predicting favorite restaurants [J]. World Wide Web, 2017, 20(6): 1313-1331.

[17] KASSAK O, KOMPAN M, BIELIKOVA M. Personalized hybrid recommendation for group of users: top-N multimedia recommender [J]. Information Processing & Management, 2016, 52(3): 459-477.

[18] LIN J, YAN X H, HUANG B. Collaborative filtering recommendation algorithm based on SVD and fuzzy clustering [J]. Computer Systems and Applications, 2016, 25(11): 156-163.

[19] ZAZOUR H, AL-SHARIF Z, AL-AYYOUB M, et al. A new collaborative filtering recommendation algorithm based on dimensionality reduction and clustering techniques [C] //Proceedings of 2018 the 9th International Conference on Information and Communication Systems. Irbid: IEEE, 2018: 102-106.

[20] 陈维亚，潘鑫，方晓平. 基于K-means聚类组合模型的公交线路客流短时预测 [J]. 华南理工大学学报(自然科学版), 2019, 47(4): 83-89.
CHEN Weiya, PAN Xin, FANG Xiaoping. Short-term prediction of passenger flow on public transit line based on K-MEANS clustering combination model [J]. Journal of South China University of Technology (Natural Science Edition), 2019, 47(4): 83-89.

[21] BAGCI H, KARAGOZ P. Context-aware friend recommendation for location based social networks using random walk [C] // Proceedings of the 25th International Conference Companion on World Wide Web. Montréal: ACM, 2016: 531-536.

[22] KONSTAS I, STATHOPOULOS V, JOSE J M. On social networks and collaborative recommendation [C] // Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Boston: ACM, 2009: 195-202.

[23] VOZALIS M G, MARGARITIS K G. Applying SVD on item-based filtering [C] // Proceedings of the 5th International Conference on Intelligent Systems Design and Applications. Warsaw: IEEE, 2005: 464-469.

[24] LI T, GAO C, DU J. A NMF-based privacy-preserving recommendation algorithm [C] // Proceedings of 2009 the First International Conference on Information Science and Engineering. Nanjing: IEEE, 2009: 754-757.

[25] LI H, MA X P, SHI J. Incorporating trust relation with PMF to enhance social network recommendation performance [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2016, 30(6): 1659016/1-28.

[26] WEIMER M, KARAZOGLOU A, BRUCH M. Maximum margin matrix factorization for code recommendation [C] // Proceedings of the Third ACM Conference on Recommender Systems. New York: ACM, 2009: 309-312.

[27] FENG X, YU W, LI Y. Faster matrix completion using randomized SVD [C] // Proceedings of 2018 IEEE the 30th International Conference on Tools with Artificial Intelligence. Volos: IEEE, 2018: 608-615.

[27] YUAN X, HAN L, QIAN S, et al. Singular value decomposition based recommendation using imputed data [J]. Knowledge-Based Systems, 2019, 163: 485-494.

Collaborative Score Prediction Method for Non-Random Missing Data

GU Wanrong[1]　*XIE Xianfen*[2]　*ZHANG Ziye*[3]　*MAO Yijun*[1]　*LIANG Zaoqing*[1]　*HE Yichen*[1]

(1. School of Mathematics and Information, South China Agricultural University, Guangzhou 510642, Guangdong, China;
2. School of Economics, Jinan University, Guangzhou 510632, Guangdong, China;
3. School of Mathematics, South China University of Technology, Guangzhou 510640, Guangdong, China)

Abstract: Most score prediction studies are based on the assumption that the missing values are random. However, the missing data of the score matrix of the actual on-line recommendation system is non-random. Incorrect assumptions about the missing data can lead to biased parameter estimation and prediction. In order to improve the accuracy of non-random missing score matrix filling, the internal principle of user and item score matrix was analyzed in this paper. It presents a method to transform the score matrix of user and object into the equivalent bilateral block diagonal matrix by row or column transformation. Then the matrix decomposition method was applied to different blocks to decompose and predict the score, making local data update and decomposition become a reality. The experimental results on the public test dataset show that the proposed method can improve the score filling effect, solve the problem of non-random score missing effectively, and improve the prediction accuracy of the recommendation system. The improved block matrix also has a better speedup ratio in the distributed processing experiment, which shows that the proposed method has better scalability.

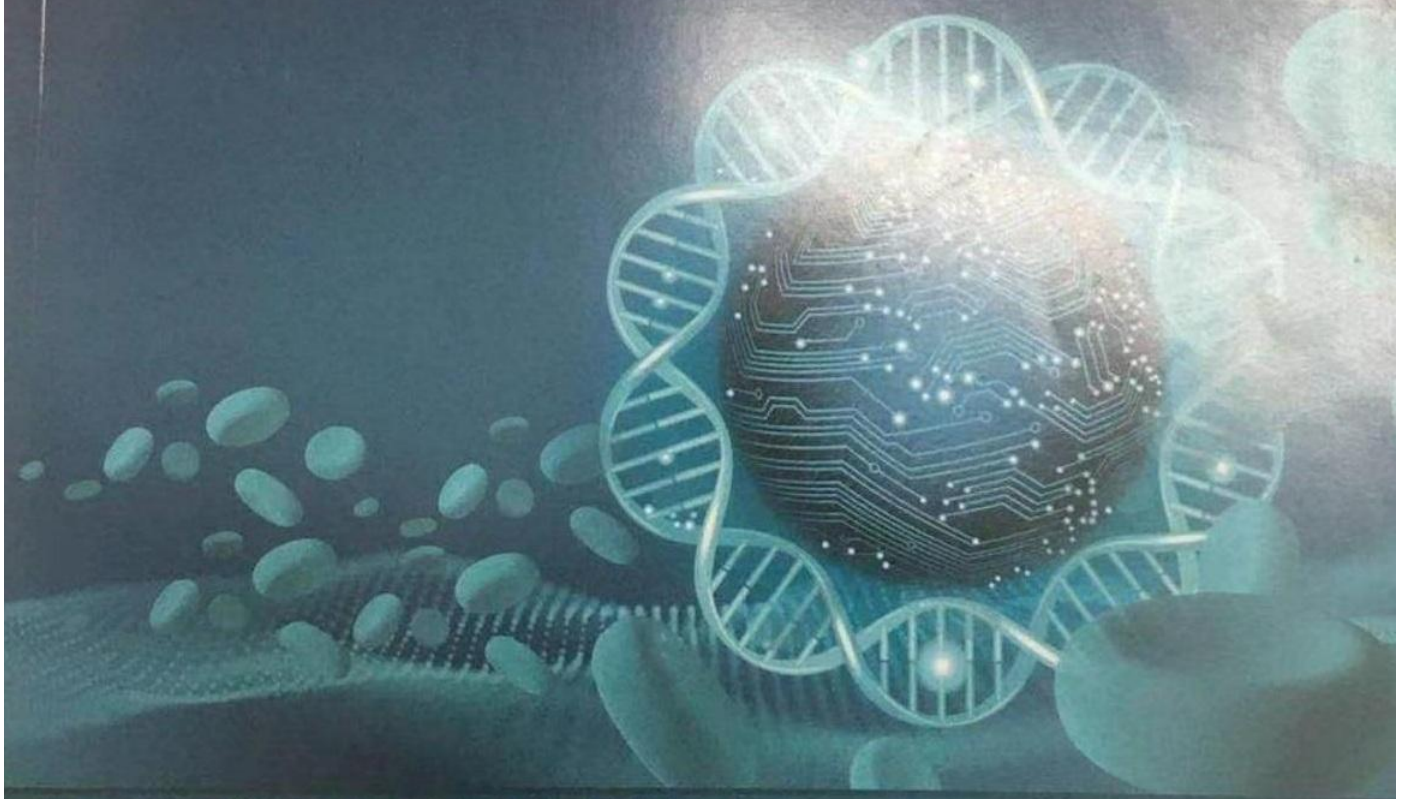
Key words: matrix decomposition; recommendation system; singular value decomposition; score prediction

ISSN 1001-5515
CN 51-1258/R
CODEN:SYGZF2

生物医学工程学杂志

SHENGWU YIXUE GONGCHENGXUE ZAZHI

JOURNAL OF BIOMEDICAL ENGINEERING



主办：四川大学华西医院　四川省生物医学工程学会

ISSN 1001-5515
9 771001 551181 12

2018年12月
第 35 卷
Vol.35 No.6
December 2018

06

生物医学工程学杂志

SHENGWU YIXUE GONGCHENGXUE ZAZHI

双月刊　1984 年 4 月创刊　第 35 卷　第 6 期（本卷终）　2018 年 12 月 25 日出版

主　管
四川省科学技术协会

主　办
四川大学华西医院
四川省生物医学工程学会

编辑出版
生物医学工程学杂志编辑部
四川省成都市武侯区国学巷 37 号，610041
电话：+86-28-85422073
Email：bjb@biomedeng.cn
网址：http://www.biomedeng.cn

荣誉主编
陈槐卿

主　编
李虹

副主编
李良，樊瑜波，Cheng Zhu，Zong-Ming Li

编辑部主任
李蓓兰

广告发布登记编号
川广登字 [2017]046 号

印　刷
成都市富生实业有限公司

公开发行
国内：四川省报刊发行局
国外：中国国际图书贸易总公司
（北京 399 信箱，100044）
代号：BM4879

订　阅
全国各地邮局
邮发代号：62-65

邮　购
生物医学工程学杂志编辑部
四川省成都市武侯区国学巷 37 号，610041
电话：+86-28-85422060

国内定价
每期 25.00 元，全年 150.00 元

中国标准连续出版物号
ISSN 1001-5515　CN 51-1258/R
CODEN SYGZF2

2018年版权归生物医学工程学杂志编辑部所有

本刊发表的文章文责由作者自负

目　次

论　著

工频电磁场长期作用影响工作记忆中局部场电位因果网络连接特征……李双燕，温雪路，桑浩钧，等　829

基于 2-back 任务下 ERP 特征的脑力疲劳客观评价研究……范晓丽，赵朝义，罗虹，等　837

面向鲤鱼机器人控脑技术的磁共振坐标转换方法研究及应用……彭勇，王婷婷，王占秋，等　845

基于样本熵和个体化阈值的表面肌肌电信号起点检测……赵翠莲，马双驰，柳叶潇　852

基于脊髓内功能电激励的大鼠后肢运动选择性控制……沈晓燕，王志功，马磊，等　860

超声介导微泡对细胞的声致成孔作用中次 Bjerknes 力的影响……徐良，林钟石，沈圆圆，等　864

基于计算机断层成像及逆向工程软件建立左心房憩室血流动力学有限元模型……文军，张林，陈国平，等　870

高强度聚焦超声脑肿瘤治疗焦域温度均匀分布调控的数值仿真研究……常诗卉，王佩国，管喜岐　877

下肢假肢接受腔对残肢肌肉萎缩影响的数值研究……王振泽，徐智，晏菲，等　887

基于先验边界条件的冠脉光学相干断层影像血管内膜轮廓序列提取……杨建利，崔国旗，李毅，等　892

靶向捕获内皮祖细胞磁性纳米粒子的生物学修饰及其细胞相容性评价……范仪，杜泽煜，黄楠　900

载穿心莲内酯胶原缓释支架促进体外炎症环境下软骨细胞表型维持的研究……徐丽女，刘海蓉，戴瑞，等　905

计算机断层扫描纹理分析在预测联合靶向化疗后的结直肠癌肝转移灶治疗反应的价值研究……张菁，周裕文，邱萌，等　914

新技术与新方法

双组分快速固化生物软组织黏合剂的研制……黄娟，李浩榕，王谦 921
基于栈式深度多项式网络集成学习框架的帕金森病计算机辅助诊断……陈璐，施俊，彭博，等 928
基于 AdaBoost 算法的药物—靶向蛋白作用预测算法……古万荣，谢贤芬，何赤琛，等 935
直接脑控多机器人协作任务研究……张超，熊馨，任弘锦，等 943
一种心电干扰下的表面肌电起止点检测方法……徐伟钊，莫鸿强，田联房，等 953

综　述

腹主动脉瘤性别差异性的生物力学研究进展……刘明，张亚星，高祖健，等 959
基于细胞的指数富集配体系统进化技术的肺癌细胞核酸适配体的筛选及其在肿瘤诊治中的应用……许金苓，廖世奇，田彩萍，等 964
生成式对抗网络在医学图像处理中的应用……潘丹，贾龙飞，曾安，等 970
卷积神经网络及其在医学图像分析中的应用研究……梁蒙蒙，周涛，张飞飞，等 977
内源性成体神经干细胞神经发生机制的研究进展……应一博，陈敏，朱娟娟，等 986

[期刊基本参数] CN 51-1258/R*1984*b*A4*164*zh*P* ￥25.00*700*23*2018-12
本期责任编辑：李蓓兰　陈咏竹

·新技术与新方法·

基于AdaBoost算法的药物—靶向蛋白作用预测算法

古万荣[1,2]，谢贤芬[2,3]，何亦琛[1]，张子烨[4]

1. 华南农业大学 数学与信息学院 计算机科学与工程系（广州 510642）
2. 广东省大数据分析与处理重点实验室（广州 510006）
3. 暨南大学 经济学院 统计系（广州 510632）
4. 华南理工大学 数学学院 数学系（广州 510660）

【摘要】 对靶向蛋白的药物作用进行预测可以促进药物新作用的发现。新近的研究更倾向于单独将特定的矩阵填补算法应用在靶向蛋白和药物的相互作用预测中。单模型的矩阵填补算法准确度较低，因此应用在药物—靶向蛋白作用预测方面也难以获得满意的结果。AdaBoost 算法是一种由多分类器组合生成强分类器的算法框架，其在分类应用领域的实用性和有效性已被证明。靶向蛋白的药物作用预测是一个矩阵填补问题，即是一种评分预测过程，因此本文在使用 AdaBoost 算法对药物—靶向蛋白作用进行预测前，将药物—靶向蛋白作用预测的矩阵填补问题转化为分类问题，将 AdaBoost 算法应用在靶向蛋白的药物作用预测评分上，充分利用 AdaBoost 算法框架对多个弱分类器进行融合从而提升性能，得以进行准确的药物—靶向蛋白作用预测。基于公测数据集的实验结果表明，本文提出的算法在预测准确度方面超过了大多数经典算法和新近算法，较好地克服了新近基于机器学习方法单算法的局限性，更好地挖掘隐含因素，有效提升了预测准确度。

【关键词】 靶向蛋白；药物作用预测；评分预测；AdaBoost 算法

Drug-target protein interaction prediction based on AdaBoost algorithm

GU Wanrong[1,2], XIE Xianfen[2,3], HE Yichen[1], ZHANG Ziye[4]

1. Department of Computer Science and Engineering, School of mathematics and information, South China Agricultural University, Guangzhou 510642, P.R.China
2. Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou 510006, P.R.China
3. Department of Statistics, School of Economy, Jinan University, Guangzhou 510632, P.R.China
4. Department of Mathematics, School of Mathematics, South China University of Technology, Guangzhou 510660, P.R.China
Corresponding author: XIE Xianfen, Email: txiexianfen2009@jnu.edu.cn

【Abstract】 The drug-target protein interaction prediction can be used for the discovery of new drug effects. Recent studies often focus on the prediction of an independent matrix filling algorithm, which apply a single algorithm to predict the drug-target protein interaction. The single-model matrix-filling algorithms have low accuracy, so it is difficult to obtain satisfactory results in the prediction of drug-target protein interaction. AdaBoost algorithm is a strong multiple classifier combination framework, which is proved by the past researches in classification applications. The drug-target interaction prediction is a matrix filling problem. Therefore, we need to adjust the matrix filling problem to a classification problem before predicting the interaction among drug-target protein. We make full use of the AdaBoost algorithm framework to integrate several weak classifiers to improve performance and make accurate prediction of drug-target protein interaction. Experimental results based on the metric datasets show that our algorithm outperforms the other state-of-the-art approaches and classical methods in accuracy. Our algorithm can overcome the limitations of the single algorithm based on machine learning method, exploit the hidden factors better and improve the accuracy of prediction effectively.

【Key words】 target protein; drug effect prediction; score prediction; AdaBoost algorithm

DOI：10.7507/1001-5515.201802026
基金项目：国家重点研发计划重点专项（2017YFC1601701）；广东省大数据分析与处理重点实验室开放基金项目（2017006，2017010）
通信作者：谢贤芬，Email：txiexianfen2009@jnu.edu.cn

Journal of Biomedical Engineering, Dec. 2018, Vol. 35, No.6

引言

在药物发现研究中，对药物—靶向蛋白相互作用（drug-target interaction，DTIs）预测是一个重要的研究热点。当前，一种新药物的发现仍然是一个非常昂贵且低效的过程，大约每一个新分子实体（new molecular entity，NME）需要花费 18 亿美元的研发成本[1]。此外，新药从研发到市场销售往往需要花费约 10 年时间。举个例子，每年只有约 20 种新药被美国食品和药品监督管理局（food and drug administration，FDA）批准。近年来，成功获批的新药数量在逐年下降。在上述情况下，对旧药物的新作用的挖掘成为一个新的研究方向。

随着药物—靶向蛋白作用研究的深入，相关的公测数据库和研究在逐年增加。例如，DrugBank 数据库具有丰富的生物和化学标注数据[2]，且具有详细的药物数据资料，包括药物类型、药物简介、化学结构、药物成分、药物靶点、药物相互作用等，该数据库由加拿大卫生研究院提供支持。此外，化学相互作用搜索工具（search tool for interactions of chemicals，STITCH）是一种预测化学—蛋白相互作用的公测数据集[3]。STITCH 数据库的关联数据主要来源于高通量实验、基因组上下文预测、共同表达、自动化文本挖掘等，更详细的信息参见本文实验部分。新近的研究表明，经 FDA 批准的药物中约 96% 标注有靶向作用蛋白。然而，仍然有许多靶向蛋白和其他药物的作用关系是未知的。

过去数十年，有许多学者对药物—靶向蛋白进行了系统研究，而传统的药物发现主要遵循“一个分子，一个目标，一种疾病”的范式，这种历史范式确定了一些影响特定蛋白质的有效化学分子。然而，这种传统的药物发现范式的主要局限性在于，药品的设计目标是针对疾病系统中的个体因素，但是复杂的疾病在本质上是多因素的，即疾病及症状是不同基因和途径的突变和干预的逐渐积累，在这种多重因素中药物发现的准确性很容易受到影响。传统模式忽视了药物与其靶向蛋白之间复杂的相互作用，而许多复杂疾病往往与多种目标蛋白有关，因此这种模式并没有像预期的那样加速药物发现的速度[4-5]。大多数药物靶点是细胞蛋白，其目的是通过与化合物有选择性的相互作用来治疗或诊断疾病。目前的研究表明，经典的治疗性药物靶标包含 130 个蛋白质家族，如酶、G 蛋白偶联受体（G-protein-coupled receptors，GPCRs）、离子通道和转运体、核激素受体等[6-7]。据估计，人类基因组中约有 6 000 ~ 8 000 个具有药理学价值的靶点，但是到目前为止，这些目标中只有小部分验证了与已批准的药物之间的相互作用，大量假定的药物靶点仍有待验证[7-9]。由于药物和蛋白的特征的高维性和复杂性，新近研究提出使用机器学习方法对药物和蛋白的相互作用进行自动挖掘，借助于数值分析和挖掘的手段，将药物和靶向蛋白作用问题转化为协同过滤、矩阵分解或评分预测问题，提高了药物和靶向蛋白作用预测的效率。虽然相对于传统药物发现，使用机器学习方法已经具备明显优势，且随着人工智能和高性能计算的发展，借助于计算机手段进行该领域的研究和应用也将成为未来的发展方向，但就目前而言，该领域内的研究仍然存在至少以下两个问题：

（1）在单个机器学习算法中，很难将复杂的参数和隐含的作用关联挖掘出来[10]；另外，在单个机器学习算法中，如果模型过于简单，则会影响预测准确率，如果过于复杂，即参数过多，则会产生过拟合问题。

（2）药物冷关联问题仍然无法在基于矩阵填补思想的预测算法中得到很好的解决[11]；冷关联问题，是指在推荐算法或评分预测研究中，存在部分对象和其他对象关联很少的现象，如果冷关联太多，对预测算法具有较大影响。

为了解决以上问题，本文提出一种较高效的基于 AdaBoost 算法的提升方法，即将多个弱分类器组合在一起，形成一个强的分类器。同时，本文通过对 AdaBoost 算法进行改进，将评分预测问题转化为分类问题，将算法框架应用到药物—靶向蛋白相互作用预测的矩阵填补研究中。本文提出的方法利用机器学习算法进行药物—靶向蛋白作用预测，相较于传统方法，具有高效、低成本的优势，同时也较好地克服了新近基于机器学习方法的单算法的局限性，能够更好地挖掘隐含因素，有效提升预测准确度，在药物冷关联问题方面，也比新近方法更好。

1　相关研究介绍

新近的药物—靶向蛋白作用预测可以大致分为 3 类：基于网络模型的方法、基于机器学习的方法以及其他模型方法。

（1）基于网络模型的方法

基于网络模型的方法是新近比较常用的药物—靶向蛋白作用预测方法。周福家等[12]提出了寻找多目标优化解决方法（multiple target optimal interven-

tion，MTOI）来获得最佳的疾病转换流程，通过最佳的疾病转换流程的挖掘来确定药物—靶向蛋白作用。Campillos 等[11]提出了一种基于药物边缘效应相似度的方法来确定两种药物是否对于同一种靶向蛋白具有相似作用。Chen 等[4]提出基于网络模型的随机游走方法来挖掘药物—靶向蛋白作用。

（2）基于机器学习的方法

最近，有研究将药物—靶向蛋白的点对标注为“正”、“负”标签样本，然后进行有监督的学习预测。这种情况下，“负”标签没有作用关联，因此不能确定“负”标签的标注。另外一种常见的基于机器学习的模型是将药物—靶向蛋白看作二分图模型，然后基于二分图进行挖掘，Yamanishi 等[13]提出了一种核回归的方法来挖掘药物—靶向蛋白的二分图。

（3）其他模型方法

其他模型方法包括：Keiser 等[14]提出了使用化学相似度来预估药物—靶向蛋白的相互作用关系；另外一种常用方法是使用基于概率矩阵分解的方法（probabilistic matrix factorization，PMF）来预测药物—靶向蛋白关系[15]。

2 理论基础原理

2.1 AdaBoost 算法

AdaBoost 算法是一个经典的自适应集成算法，其基本原理在于通过对弱可学习的算法进行集成，进而生成强可学习的模型。该算法需要有输入数据集 $\boldsymbol{D}$，若干个基础分类模型集合 $\boldsymbol{C}$。该算法首先通过对若干个训练样本的学习训练获得第一个弱分类器；然后，将错误分类的样本和其他未训练样本一起构成新的训练样本集，通过再次训练获得第二个弱分类器；再然后，将以上两次学习都分类错误的样本加上未学习的样本一起构成新的训练样本，进而得到第三个弱分类器；以此类推，通过若干次迭代后，该分类器得到性能加强和提升，逐渐变成强分类器[10]。现在，AdaBoost 算法已被 Schapire[16]证明弱可学习和强可学习是等价的。

2.2 矩阵分解模型

矩阵分解模型起源于代数中的奇异值分解，是挖掘具有二分图结构的推荐问题的重要工具之一，它通过矩阵降维的方式将原本较稀疏的矩阵缺失数据进行补全。在挖掘二分图结构关系中，可以将药物—靶向蛋白的作用关系矩阵分解为两个较低维度的矩阵的乘积，如式（1）所示：

$$\hat{\boldsymbol{R}}=\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{P} \tag{1}$$

其中，$\boldsymbol{P}\in\mathbb{R}^{k\times m}$，$\boldsymbol{Q}\in\mathbb{R}^{k\times n}$ 是两个降维后的矩阵。$\boldsymbol{P}$ 矩阵包含了全部药物及其隐含属性，$\boldsymbol{Q}$ 矩阵包含了全部靶向蛋白及其隐含属性。k 是隐含因子的维度，大小可根据经验设定。其物理意义是：假设每个药物都有 k 个隐含的属性，这些属性可以和靶向蛋白的 k 隐含属性相对应，使之产生作用关联。因此，药物对靶向蛋白的作用可以用 $\hat{\boldsymbol{R}}(d,t)=\hat{r}_{dt}$ 表示，如式（2）所示：

$$\hat{r}_{dt}\approx\sum_{k=1}^{n}p_{dk}q_{tk} \tag{2}$$

其中，$p_{dk}=\boldsymbol{P}(d,k)$，$q_{tk}=\boldsymbol{Q}(t,k)$ 在现实场景中，可能会面临更复杂的情况，因此应该对模型进行适当的修正，加入偏差系数等。

3 基于 AdaBoost 算法的药物—靶向蛋白作用预测算法

在推荐系统的评分预测模型中，一般先将用户集合和物品集合分别定义为：$\boldsymbol{U}=\{u_1,u_2,\cdots,u_{|U|}\}$，$\boldsymbol{I}=\{i_1,i_2,\cdots,i_{|I|}\}$。并以此为基础定义评分矩阵，即：$\boldsymbol{D}=\{(u,i,r_{ui})_1,(u,i,r_{ui})_2,\cdots,(u,i,r_{ui})_m\}$，共计 m 个评分，一般而言，有效的评分数 $m\ll|\boldsymbol{U}|\times|\boldsymbol{I}|$。

然而，传统的 AdaBoost 算法框架适用于分类问题，如前文的算法框架伪代码。如要将其应用到药物—靶向蛋白作用预测中，则需要将矩阵分解模型转化为分类模型，因此需要对 AdaBoost 算法模型进行改进，分别需要解决以下问题：

（1）如何将评分预测问题转化为分类问题，即将药物—靶向蛋白作用预测的二值矩阵分解和填补问题转化为 AdaBoost 算法框架可以应用的问题。

（2）模型错误率的评估和计算。

（3）训练数据的分类权重分布如何与矩阵分解模型结合。

（4）如何将多个模型集成起来。

3.1 问题转化

AdaBoost 算法对多弱分类模型集成的性能已经得到较多研究成果确认。因此，本文将药物—靶向蛋白作用预测问题转化为二值分类问题。首先，引入一个参数 ϕ 作为阈值，该阈值的作用是用于评判矩阵分解模型中的评分预测结果是好、还是坏。假设某药物对某靶向蛋白的作用预测结果为 3，通过矩阵分解模型得出的分值如果在$[3-\phi,3+\phi]$区间内，则认为该矩阵分解模型是比较准确的，反之则认为该模型不准确。通过这种映射，将矩阵分解的评分预测模型转化为二值分类问题。

3.2 模型错误率评估

分类方法的模型错误率很容易通过错误分类样本进行统计，评分预测问题的错误率评估往往通过预测结果 $\hat{r}_{dt}$ 和真实结果 r_{dt} 的差距来表示，比如以均方根误差（root mean square error，RMSE）或平均绝对误差（mean absolute error，MAE）等来表示。在本文提出的方法处理中，当预测结果 $\hat{r}_{dt} \in [r_{dt}-\phi, r_{dt}+\phi]$ 时，被认为是准确的，即不计算其错误率，否则使用以下计算，如式（3）所示：

$$\varepsilon_i = \sum_{d=1}^{|\boldsymbol{d}|}\sum_{t=1}^{|\boldsymbol{t}|} w_{dt} \tag{3}$$

公式中，$\boldsymbol{d}$ 表示药物集合，$\boldsymbol{t}$ 表示靶向蛋白集合。

3.3 权重更新

AdaBoost 算法的有效性很大程度上依赖于各种基础模型的权重分布，即对错误的分类数据会加大权重，在下一轮迭代中会更加重视其训练，反之则降低权重，不断迭代后，分类器逐渐更具鲁棒性。最后将所有模型集成起来以达到最优效果。因此，本文定义损失函数如式（4）所示：

$$\arg\min \sum_{d=1}^{|\boldsymbol{d}|}\sum_{t=1}^{|\boldsymbol{t}|} (r_{dt}-\hat{r}_{dt})^2 \tag{4}$$

将损失函数加入到每个训练数据中以增强训练效果，如式（5）所示：

$$\arg\min \sum_{d=1}^{|\boldsymbol{d}|}\sum_{t=1}^{|\boldsymbol{t}|} w_{dt}(r_{dt}-\hat{r}_{dt})^2 \tag{5}$$

本文使用随机梯度下降法来实现药物—靶向蛋白作用矩阵中的预测问题，即将权重赋值给如式（6）所示：

$$\arg\min \sum_{d=1}^{|\boldsymbol{d}|}\sum_{t=1}^{|\boldsymbol{t}|} w_{dt}[(r_{dt}-\mu-\boldsymbol{b}_t-\boldsymbol{b}_d-\boldsymbol{q}_t{}^{\mathrm{T}}\boldsymbol{p}_u)^2 + \lambda(\boldsymbol{b}_t{}^2+\boldsymbol{b}_d{}^2+||\boldsymbol{q}_i||^2+||\boldsymbol{p}_u||^2)] \tag{6}$$

其中，w_{dt} 是药物 d 对于靶向蛋白 t 的作用评分的权重。每个因子的递推迭代公式如式（7）所示：

$$\begin{aligned} \boldsymbol{b}_t &\leftarrow \boldsymbol{b}_t + w_{dt}\cdot\gamma\cdot(e_{dt}-\lambda\cdot\boldsymbol{b}_t) \\ \boldsymbol{b}_d &\leftarrow \boldsymbol{b}_d + w_{dt}\cdot\gamma\cdot(e_{dt}-\lambda\cdot\boldsymbol{b}_d) \\ \boldsymbol{q}_t &\leftarrow \boldsymbol{q}_t + w_{dt}\cdot\gamma\cdot(e_{dt}\boldsymbol{p}_u-\lambda\cdot\boldsymbol{q}_t) \\ \boldsymbol{p}_u &\leftarrow \boldsymbol{p}_u + w_{dt}\cdot\gamma\cdot(e_{dt}\boldsymbol{q}_t-\lambda\cdot\boldsymbol{p}_u) \end{aligned} \tag{7}$$

根据前面的错误率，权重更新公式如式（8）所示：

$$\alpha_i = \mu\ln\left(\frac{1-\varepsilon_i}{\varepsilon_i}\right) \tag{8}$$

其中，参数 μ 是调整参数，用于自适应训练数据的调整。

3.4 模型集成

通过 AdaBoost 算法展开训练，可以集成 T 个基础训练模型的融合模型，并通过加权平均的方法来集成各个模型的药物—靶向蛋白作用预测框架，每个模型有一个对应权重 α_i，最终的药物—靶向蛋白作用预测结果如式（9）所示：

$$\hat{\boldsymbol{R}}_{dt}(d,t) = \sum_{i=1}^{T}\alpha_i h_i(d,t) \tag{9}$$

3.5 算法设计与分析

基于前文讨论，AdaBoost 算法框架的药物—靶向蛋白作用预测算法具体如下：

输入：数据集 $\boldsymbol{D}=\{(d,t,r_{dt},w_{dt})_1,(d,t,r_{dt},w_{dt})_2,\cdots,(d,t,r_{dt},w_{dt})_m\}$；基础评分预测模型 L，如奇异值分解方法（singular value decomposition，SVD）或概率矩阵分解（probabilistic matrix factorization，PMF）等；训练轮数 T；更新权重的控制参数 μ；

输出：$\hat{r}_{dt}(d,t) = \sum_{i=1}^{T}\alpha_i h_i(d,t)$

1：$(d,t,r_{dt},w_{dt})_t^1 = \frac{1}{m}$；

2：**for** $i=1,\cdots,T$ **do**

3：$h_i = L(\boldsymbol{D},\boldsymbol{D}_i)$;

4：**for** all $\hat{r}_{dt} \notin [r_{dt}-\phi, r_{dt}+\phi]$ in $\boldsymbol{D}_i$, the set is $\boldsymbol{D}_i^-$ **do**

5：$\varepsilon_i = \sum_{d,t\in D_i^-} w_{dt}^t$;

6：**end for**

7：$\alpha_i = \ln\left(\frac{1-\varepsilon_i}{\varepsilon_i}\right)$;

8：$\boldsymbol{D}_{i+1}(t) = \frac{\boldsymbol{D}_i(t)}{Z_i}\times\begin{cases}\exp(\alpha_i)\ if\ h_i(x_t)\neq y_t \\ \exp(-\alpha_i)\ if\ h_i(x_t)=y_t\end{cases}$;

9：**end for**

上述算法有明确的输入，即需要基本的算法模型，如 SVD 奇异值分解算法；其次，需要有训练集 $\boldsymbol{D}$、训练轮数 T、判断作用分准确与否的阈值 ϕ 以及更新权重的控制参数 μ。

4 实验与结果分析

本文实验基于两个公测数据集：人工标注靶点和药物在线资源（manually annotated targets and drugs online resource，MATADOR）和前文“引言”部分提及的 STITCH 数据集[3, 17]，对比分析了方法中的参数及对预测结果的影响，并针对若干经典分类方法和新近方法进行仿真分析。

4.1 数据集与评测指标

本文的实验数据集主要来自两个公开数据集：① MATADOR（网址为：http://matador.embl.de/）[17]，是一个公开的来源于人工整理和自动文本挖掘的在线治疗数据库，其中包含了许多药物—靶向蛋白的相互作用关系数据。本文实验选用了该数据库中具有药物—靶向蛋白直接或间接作用关系的数据。MATADOR 数据库中包含了 784 种药物，2 431 种靶向蛋白和 13 064 个药物—靶向蛋白相互作用，其中共包含了 7 862 种直接和 5 202 种间接作用关系。② STITCH（网址为：http://stitch.embl.de/）[3]，该数据库提供了许多靶向蛋白及其化学成分的相互关系数据，这些标注关系来源于已知的实验和文献，具有较高的可信度。本文选用了 STITCH 数据库中和 MATADOR 数据库重合的药物—靶向蛋白数据。STITCH 数据库中包含了 598 种药物，671 种靶向蛋白和 3 296 个有效的药物—靶向蛋白作用对，在其中分别有 2 589、945 和 1 493 条相互作用关系，分别对应于绑定、激活和隐含关系。为了能在作用预测中评分更直观，本文将数据集中的作用分数均匀映射到[0, 5]之间。

本文实验将上述数据集平均分为 5 份，然后使用 5 折交叉验证方法进行实验，其中 1 份作为测试集，剩余 4 份作为训练集。由于药物—靶向蛋白作用预测是一种评分预测问题，因此常使用 RMSE（以符号 *RMSE* 表示）方法作为评价指标[18]，定义为预测值与真实值之间的差的平方均值，如式（10）所示：

$$RMSE = \frac{\sqrt{\sum_{d=1}^{|d|}\sum_{t=1}^{|t|}(r_{dt}-\hat{r}_{dt})^2}}{|\boldsymbol{D}|} \tag{10}$$

与 RMSE 评测标准同样常用的是 MAE，区别在于 RMSE 对于错误作用预测的惩罚扣分会更多，即对错误结果更苛刻。

本文还使用了预测准确率（accuracy，Accu）和召回率（recall，Rec）两个指标。这两个指标的计算依赖于实际样本数据和预测统计数据，如表 1 所示，真阳性（true positive，TP）（以符号 *TP* 表示），指具体的药物—靶向蛋白具备相互作用且预测结果显示也具备相互作用的统计结果总数。以此类推，假阴性（false negative，FN）（以符号 *FN* 表示），指药物—靶向蛋白具备相互作用，但预测结果显示无相互作用的统计结果总数；假阳性（false positive，FP）（以符号 *FP* 表示），指药物—靶向蛋白不具备相互作用，但预测结果显示有相互作用的统计结果总数；真阴性（true negative，TN）（以符号 *TN* 表示），指药物—靶向蛋白不具备相互作用，预测结果也显示无相互作用的统计结果总数。

通过以上的统计数据，Accu（以符号 *Accu* 表示）的计算结果如式（11）所示：

$$Accu = (TP + TN) / (TP + TN + FN + FP) \tag{11}$$

Rec（以符号 *Rec* 表示）的计算结果如式（12）所示：

$$Rec = TP / (TP + FN) \tag{12}$$

4.2 实验结果与分析

矩阵分解模型是评分预测模型中常用的方法，因此，本文的 AdaBoost 算法框架使用 SVD 方法和 PMF 模型作为药物—靶向蛋白作用预测的基本预测模型。矩阵分解过程中的学习速率和惩罚系数是重要的预设参数，根据模型训练经验，本文在 SVD 方法中设定的学习速率为 0.01、惩罚系数为 0.01，对于 PMF 模型，设定学习速率为 0.005、惩罚系数为 0.02。

4.2.1 AdaBoost 算法框架对比分析　如表 2 所示，展示了本文提出的 AdaBoost 算法在药物—靶向蛋白中的作用预测方法。从实验结果可知，使用了 AdaBoost 算法相对于单个模型而言，效果有了较大的提升。

从实验结果可知，在两个不同的数据集和不同的基础模型中，使用了 AdaBoost 算法的运算结果的误差较小。其原因在于，使用 AdaBoost 算法将基础方法进行融合提升，通过对分类错误进行二次迭代后不断逼近，以此达到预测准确度提升的预期效果。

4.2.2 算法框架的相关参数分析　本文所提方法的参数主要有阈值 ϕ，训练轮数 T，权重更新 μ，本实

表 1　药物—靶向蛋白作用预测的统计结果

Tab.1　Statistical results of drug-target protein interaction prediction

	预测有作用	预测无作用
实际有作用	*TP*	*FN*
实际无作用	*FP*	*TN*

表 2　AdaBoost 算法实验效果

Tab.2　Experimental results of AdaBoost algorithm

算法	RMSE			
	MATADOR		STITCH	
	未使用	AdaBoost	未使用	AdaBoost
SVD	0.923 4	**0.892 2**	0.832 1	**0.807 6**
PMF	0.943 8	**0.912 3**	0.859 8	**0.832 1**

验基于 STITCH 数据集，使用 SVD 方法进行实验，实验结果如图 1 所示。

从以上实验结果可知，阈值 ϕ 的选取对 RMSE 实验结果的影响是先大再小，然后再变大，阈值 ϕ 在[0.1，1.0]取值的 10 个实验数据中，当取 0.7 时，使得 RMSE 值最小。所以，如果想要得到最好的预测准确度，应当调整适当的阈值 ϕ，过大或过小都会对预测造成不良影响。

如图 2 所示，随着训练轮数的增长，一开始精度提高较快，达到 7 次以上时，精度略有提高，但趋势变缓。

如图 3 所示，权重更新参数 μ 对药物—靶向蛋白作用预测也具有明显的影响，从实验可知，μ = 0.4 时，准确度最高。

4.2.3 *其他方法的准确度对比与分析*　本实验实现了一些经典的评分预测方法和新近方法作为基线方法进行对比，具体如下：

基于近邻用户的推荐算法（user-based）[19]：该算法应用到药物—靶向蛋白预测中，将近邻药物看作是近邻用户。

基于近邻物品的推荐算法（item-based）[20]：该算法应用到药物—靶向蛋白预测中，将近邻蛋白看作是近邻物品。

朴素贝叶斯算法（Naïve Bayes）[21]：是一种基于贝叶斯后验概率的分类预测方法，常用于数据分类和评分预测。

支持向量机（support vector machine，SVM）方法[22]：这是一种性能比较优越的分类器模型，该算法使用超平面进行分类训练，对非线性数据也可以采用核技术将其进行准确分类[23]。

MTOI（周福家等[12]提出）：通过寻找多目标优化解决方法来获得最优的疾病转换状态，通过最优的疾病转换状态的挖掘来确定药物—靶向蛋白作用。

基于二分图的核回归方法（Yamanishi 等[13]提出）：该方法通过将药物—靶向蛋白关系映射为二分图，然后通过核回归方法挖掘二分图的待预测节点。

基于化学相似度的方法（Keiser 等[14]提出）：该方法通过蛋白质的化学相似度作为挖掘基础，从而推断药物—靶向蛋白的作用关系。

以上方法的最终实验结果如表 3 所示，“本文方法”是指本文使用 AdaBoost 算法的结果，基础模型使用了 SVD 方法。

从表3实验结果可知，本文算法获得了较好结果，其原因在于，本文使用 AdaBoost 算法将基础方法如 SVD 方法进行迭代融合，多轮迭代更逼近准确结果。从表 2 也可以看出，即便 SVD 方法不使用模型融合也获得不差的结果，在使用了 AdaBoost 算法后，准确度明显上升。

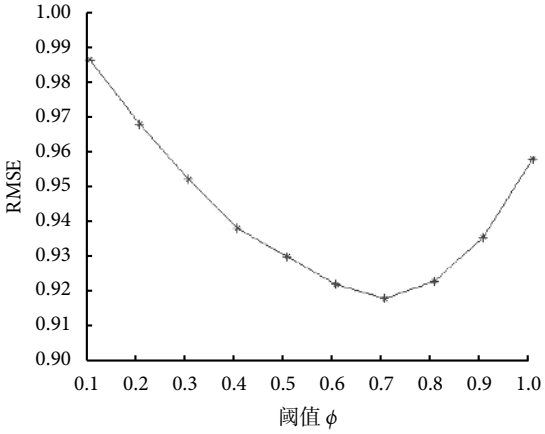


图 1　阈值 ϕ 对预测结果的影响

Fig.1　The influence of threshold ϕ

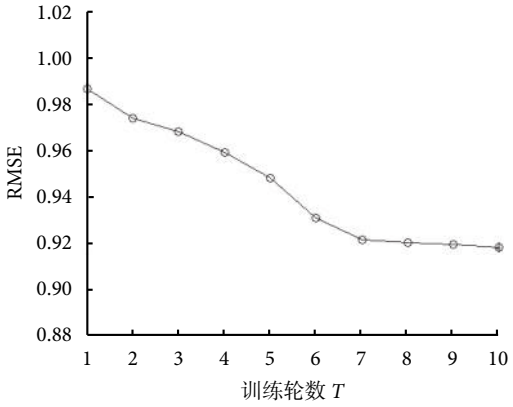


图 2　训练轮数 T 对预测结果的影响

Fig.2　The influence of training iterations T

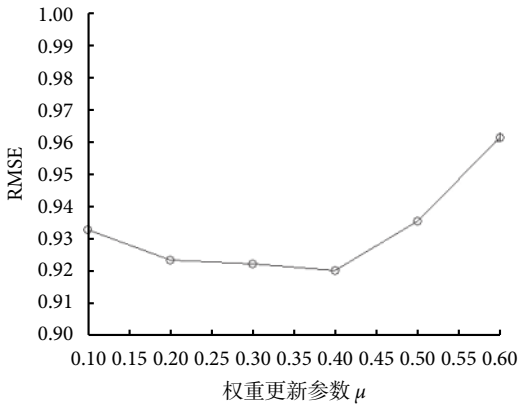


图 3　权重更新参数 μ 对预测结果的影响

Fig.3　The influence of weight updating parameter μ

4.2.4 *冷关联问题的实验*　在药物—靶向蛋白作用预测中，本文发现在训练集中，有些药物关联的蛋

表 3 本文方法与新近方法的实验结果对比

Tab.3 Comparison of recent approaches and our proposed method

算法	数据集	Accu	Rec	RMSE
user-based 方法	MATADOR	89.12%	85.12%	1.115
	STITCH	88.54%	84.62%	1.295
item-based 方法	MATADOR	88.45%	84.12%	1.138
	STITCH	86.41%	82.89%	1.345
Naïve Bayes 方法	MATADOR	91.52%	88.45%	1.038
	STITCH	89.92%	86.95%	1.120
SVM 方法	MATADOR	91.92%	89.21%	0.914
	STITCH	90.10%	88.32%	0.983
MTOI 方法	MATADOR	91.35%	90.35%	0.902
	STITCH	90.65%	89.65%	0.972
基于二分图的核回归方法	MATADOR	92.36%	91.62%	0.926
	STITCH	91.69%	90.62%	0.982
基于化学相似度的方法	MATADOR	90.18%	89.12%	0.933
	STITCH	88.64%	85.64%	0.950
本文方法	MATADOR	**95.10%**	**93.21%**	**0.807**
	STITCH	**94.58%**	**91.58%**	**0.892**

表 4 本文方法与新近方法在冷关联数据集上的实验结果对比

Tab.4 Comparison of recent approaches and our proposed method in cold relevance data set

算法	数据集	Accu	Rec	RMSE
user-based 方法	MATADOR-C	23.38%	20.56%	2.358
	STITCH-C	22.78%	19.78%	2.669
item-based 方法	MATADOR-C	22.89%	21.65%	2.610
	STITCH-C	20.36%	19.35%	2.721
Naïve Bayes 方法	MATADOR-C	25.45%	24.36%	2.221
	STITCH-C	19.96%	20.39%	2.389
SVM 方法	MATADOR-C	25.92%	26.75%	2.003
	STITCH-C	20.17%	**25.96%**	2.169
MTOI 方法	MATADOR-C	25.65%	24.98%	1.985
	STITCH-C	24.89%	23.54%	2.069
基于二分图的核回归方法	MATADOR-C	27.11%	26.52%	1.845
	STITCH-C	26.05%	24.85%	1.896
基于化学相似度的方法	MATADOR-C	26.25%	25.59%	1.956
	STITCH-C	24.47%	22.84%	2.056
本文方法	MATADOR-C	**29.02%**	**26.87%**	**1.782**
	STITCH-C	**27.54%**	24.87%	**1.806**

白非常少，这就是评分预测研究中常见的“冷关联问题”，冷关联问题加剧了数据挖掘中的数据稀疏性的负面影响。本文对实验选取的数据集进一步整理，形成关联较少的数据集，以验证本文方法对冷关联数据集下的有效性。首先，对 MATADOR 数据集和 STITCH 数据集进行整理，挑选了药物关联数少于 3 的药物、蛋白以及关联信息。经过挑选后，MATADOR 数据库中包含了 365 种药物，958 种靶向蛋白和 5 211 个有效的药物—靶向蛋白作用对，其中共包含了 3 851 种直接和 1 860 种间接作用关系。STITCH 数据库中包含了 231 种药物，158 种靶向蛋白和 1 213 个有效的药物—靶向蛋白作用对，在其中分别有 512、254 和 447 条相互作用关系，分别对应于绑定、激活和隐含关系。为了便于区分，挑选后的数据库分别称为 MATADOR—冷关联（cold）（以符号 MATADOR-C 表示）和 STITCH—冷关联（cold）（以符号 STITCH-C 表示）。实验结果如表 4 所示。

从实验结果可知，冷关联数据集中的实验结果普遍都不好，不管是 Accu、Rec 还是 RMSE，各种方法都受到很大影响。本文方法除了在 STITCH-C 数据集上的 Rec 指标中，比 SVM 方法较差外，其余实验结果均获得较好结果，但该指标中的结果数据相差也较微小。

5 结语

本文提出使用融合学习模型的 AdaBoost 算法来提高药物—靶向蛋白作用预测的效果，首次在该领域使用这种融合算法。实验结果表明，本文方法具有较高的预测准确度。使用 AdaBoost 算法可以让弱分类器融合组成较强分类方法，提升分类性能，使用本文方法还能将作用预测矩阵切换为分类算法框架。本文工作是对目前新近算法常使用单机器学习模型进行药物—靶向蛋白作用预测的一种多模型融合尝试。将来的研究工作中，可能还会面临两个方面的问题函待解决：① 隐含关联特征如何进行有效提取及其如何确定其在相互作用挖掘中的权重。对于数据挖掘中进行特征工程研究，业界已经有了深度学习方法（deep learning，DL）及其相关的软硬件平台体系，并在人工智能等领域取得了良好的应用效果。如何将 DL 应用在药物—靶向蛋白作用预测将是未来的一个重要研究方向。② 海量数据的快速处理。随着药物和病种亚型的增加，相互作用关联关系迅速增长，如何处理海量的关联关系也将是今后的重要研究方向。映射—降解（map-reduce，MR）处理框架是目前流行的海量数据批处理框架，在未来工作中，为更好地处理海量的药物—靶向蛋白、蛋白—蛋白等作用预测，本研究还将持续研究本文算法在 MR 框架中的分布式集群处理方法，使其可快速获得分析结果，并推进该研究的应用实践。在特征挖掘的研究方面，本研究将使用 DL 方法，深入挖掘隐含特征集的关系，以提升药物—靶向蛋白作用的预测准确度。

参考文献

1 Paul S M, Mytelka D S, Dunwiddie C T, *et al*. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov, 2010, 9(3): 203-214.

2 Law V, Knox C, Djoumbou Y, *et al*. DrugBank 4.0: shedding new

light on drug metabolism. Nucleic Acids Research, 2014, 42(Database issue): D1091-D1097.
3 Kuhn M, Szklarczyk D, Pletscher-Frankild S A, *et al*. STITCH 4: integration of protein-chemical interactions with user data. Nucleic Acids Res, 2014, 42(D1): D401-D407.
4 Chen Xing, Liu Mingxi, Yan G Y. Drug-target interaction prediction by random walk on the heterogeneous network. Mol Biosyst, 2012, 8(7): 1970-1978.
5 Iskar M, Zeller G, Zhao Xingming, *et al*. Drug discovery in the age of systems biology: the rise of computational approaches for data integration. Current opinion in biotechnology, 2012, 23(4): 609-616.
6 Hopkins A L, Groom C R. The druggable genome. Nature reviews Drug discovery, 2002, 1: 727-730.
7 Drews J. Drug discovery: A historical perspective. Science, 2000, 287(5460): 1960-1964.
8 Overington J P, Al-Lazikani B, Hopkins A L. How many drug targets are there? Nature Reviews Drug discovery, 2006, 5: 993-996.
9 Landry Y, Gies J P. Drugs and their molecular targets: an updated overview. Fundamental & clinical pharmacology, 2008, 22(1): 1-18.
10 Zhu Ji, Zou Hui, Rosset S, *et al*. Multi-class AdaBoost. Stat Interface, 2009, 2(3): 349-360.
11 Campillos M, Kuhn M, Gavin A C, *et al*. Drug target identification using side-effect similarity. Science, 2008, 321(5886): 263-266.
12 周福家, 张宏伟, 李卫国. 分子网络多靶标筛选的粒子群数值模拟法. 计算力学学报, 2015, 32(2): 269-273.
13 Yamanishi Y, Araki M, Gutteridge A A, *et al*. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics, 2008, 24(13): I232-I240.
14 Keiser M J, Setola V, Irwin J J, *et al*. Predicting new molecular targets for known drugs. Nature, 2009, 462(7270): 175-181.
15 Cobanoglu M C, Liu Chang, Hu Feizhuo, *et al*. Predicting drug-target interactions using probabilistic matrix factorization. J Chem Inf Model, 2013, 53(12): 3399-3409.
16 Schapire R E. The strength of weak learnability. Mach Learn, 1990, 5(2): 197-227.
17 Bleakley K, Yamanishi Y. Supervised prediction of drug-target interactions using bipartite local models. Bioinformatics, 2009, 25(18): 2397-2403.
18 Takács D, Pilászy I, Németh B. Major components of the gravity recommendation system. ACM SIGKDD Explorations Newsletter, 2007, 9(2): 80-83.
19 Wu Zhenhua, Chen Xiaosu, Xiao Daoju. Offline Chinese signature verification based on segmentation and RBFNN classifier. Acta Automatica Sinica, 2007, 345(1): 995-1001.
20 Li Dongsheng, Chen Chao, Lv Qin, *et al*. An algorithm for efficient privacy-preserving item-based collaborative filtering. Future Generation Computer Systems, 2016, 55(C): 311-320.
21 Bilal M, Israr H, Shahid M, *et al*. Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques. Journal of King Saud University-Computer and Information Sciences, 2016, 28(3): 330-344.
22 Patel T B, Patil H A. Cochlear filter and instantaneous frequency based features for spoofed speech detection. IEEE J Sel Top Signal Process, 2017, 11(4): 618-631.
23 刘晓峰, 张雪英, Wang Z J. Logistic核函数及其在语音识别中的应用. 华南理工大学学报:自然科学版, 2015, 43(5): 100-106.

收稿日期：2018-02-10　修回日期：2018-07-03
本文编辑：陈咏竹

(上接第934页；*Continued from Page* 934)

4 Salvatore C, Cerasa A, Castiglioni I, *et al*. Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and Progressive Supranuclear Palsy. J Neurosci Methods, 2014, 222: 230-237.
5 Rana B, Juneja A, Saxena M, *et al*. Regions-of-interest based automated diagnosis of Parkinson's disease using T1-weighted MRI. Expert Syst Appl, 2015, 42(9): 4506-4516.
6 Adeli E, Shi Feng, An Le, *et al*. Joint feature-sample selection and robust diagnosis of Parkinson's disease from MRI data. Neuroimage, 2016, 141: 206-219.
7 Peng Bo, Wang Suhong, Zhou Zhiyong, *et al*. A multilevel-ROI-features-based machine learning method for detection of morphometric biomarkers in Parkinson's disease. Neurosci Lett, 2017, 651: 88-94.
8 Shen D, Wu G, Suk H I. Deep learning in medical image analysis. Annu Rev Biomed Eng, 2017, 19: 221.
9 Plis S M, Hjelm D R, Salakhutdinov R, *et al*. Deep learning for neuroimaging: a validation study. Front Neurosci, 2014, 8(8): 229.
10 陈诗慧, 刘维湘, 秦璟, 等. 基于深度学习和医学图像的癌症计算机辅助诊断研究进展. 生物医学工程学杂志, 2017, 34(2): 314-319.
11 Payan A, Montana G. Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. Computer Science, 2015. arXiv: 1502.02506.
12 Suk H I, Lee S W, Shen Dinggang, *et al*. Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. Brain Struct Funct, 2015, 220(2): 841-859.
13 Livni R, Shalev-Shwartz S, Shamir O. An algorithm for training polynomial networks. Computer Science, 2013, 26(18): 4748-4750.
14 Shi J, Zheng X, Li Y, *et al*. Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. IEEE Journal of Biomedical & Health Informatics, 2018, 22(1): 173-183.
15 Shi Jun, Zhou Shichong, Liu Xiao, *et al*. Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset. Neurocomputing, 2016, 194: 87-94.
16 Zhu P, Zhang L, Hu Q, *et al*. Multi-scale patch based collaborative representation for face recognition with margin distribution optimization// European Conference on Computer Vision. 2012, 7572: 822-835.
17 Yang F, Lu H, Yang M H. Robust visual tracking via multiple kernel boosting with affinity constraints. IEEE Transactions on Circuits and Systems for Video Technology, 2014, 24(2): 242-254.
18 Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell, 2013, 35(8): 1798-1828.

收稿日期：2017-09-14　修回日期：2018-02-28
本文编辑：李蓓兰

Efficient Multi-View Clustering via Essential Tensorized Bipartite Graph Learning

Wanrong Gu, Junlong Guo, Haiyan Wang, Guangyu Zhang, Bin Zhang, Jiazhou Chen, *Member, IEEE*, and Hongmin Cai, *Senior Member, IEEE*

***Abstract*—Multi-view spectral clustering has garnered significant attention for its capacity to integrate intrinsic feature information from multiple perspectives, resulting in improved performance. However, the oversight of inter-view correlations has led to suboptimal outcomes. Furthermore, the conventional method of constructing an $N \times N$ graph in multi-view clustering imposes a substantial time burden when dealing with large-scale scenarios. To address these challenges, this paper presents an efficient multi-view clustering approach via *E*ssential *T*ensorized *B*ipartite *G*raph *L*earning (ETBGL). Specifically, ETBGL utilizes the low-rank tensor Schatten p-norm to capture inter-view similarity, effectively capturing high-order correlation information embedded in multiple views. Simultaneously, by incorporating bipartite graph learning, ETBGL efficiently mitigates the computational demands and spatial complexity associated with tensor operations. Additionally, we introduce the $\ell_{2,1}$-norm of tensor as a sparse penalty to the error term, with the aim of filtering out noise and preserving shared information, thus enhancing clustering robustness. We solve our objective by an efficient algorithm that is time-economical and has good convergence. Comprehensive evaluations on diverse datasets demonstrate the exceptional performance of our proposed model.**

***Index Terms*—Bipartite graph learning, multi-view clustering, tensor schatten p-norm.**

I. Introduction

MULTI-VIEW clustering is an advanced data clustering method that leverages information from multiple views to enhance clustering performance. By integrating diverse feature information from different views, multi-view clustering can accurately reveal the intrinsic structure and similarity of data. It is particularly effective in handling heterogeneous datasets and offers strong interpretability and visualization capabilities [1], [2], [3], [4], [5]. Spectral clustering utilizes the eigenvectors of the affinity matrix to project data into a low-dimensional space for clustering, offering advantages such as handling irregularly shaped data and being insensitive to input space transformations [6], [7], [8]. Therefore, spectral clustering has gained widespread attention.

While multi-view clustering (MVC) has received significant attention, it still faces several challenges: **(1)**Traditional multi-view clustering often involves clustering individual views separately and then merging the results. This simplistic combination approach fails to fully exploit the correlations and complementarity among different views, which means that the embedded information in the data may not be fully utilized. **(2)** In multi-view clustering, computing the similarity matrix and Laplacian matrix decomposition for different views can be computationally intensive. The complexity for constructing these matrices is $O(Vn^2)$ and $O(n^3)$, respectively, where V and n represent the number of views and the number of samples. Consequently, processing large-scale datasets becomes inefficient. Additionally, when dealing with high-dimensional data, traditional multi-view clustering algorithms may be impacted by the data dimensionality, leading to significantly increased computation and storage costs.

The multi-view clustering method based on a bipartite graph has garnered significant attention [14]. This method involves partitioning the data points from each view in the multi-view dataset into two distinct sets of vertices, forming a bipartite graph. By connecting vertices from different views within this graph, it facilitates the fusion of information from various perspectives, enabling the comprehensive utilization of multi-view data. The bipartite graph comprises n data points and m anchor points, where m is considerably smaller than n ($m \ll n$), effectively reducing computational and storage complexities due to its characteristic features.

Drawing inspiration from this, D. Cai et al. [9] proposed a similar bipartite graph-based method, where representative data points are selected as anchor points and related to the original data. However, the selection of anchor points relies on empirical settings. Z. Kang et al. [10] attempted linear fusion of the learned bipartite graph, using singular value decomposition on the matrix to obtain results. Nonetheless, they did not learn genuine coefficient representations and only approximated using kernel methods. X. Li et al. [11] achieved a consensus bipartite graph through the linear combination of learned bipartite graphs. However, simple combinations fail to capture the associations

Received 18 April 2024; revised 15 August 2024; accepted 6 October 2024. Date of publication 9 December 2024; date of current version 24 July 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62202176 and Grant 62206099, in part by the Natural Science Foundation of Guangdong Province under Grant 2023A1515012885 and Grant 2022A1515011489, and in part by the National Statistical Science Research Project under Grant 2023LY016. *(Corresponding author: Haiyan Wang.)*

Wanrong Gu, Junlong Guo, Haiyan Wang, and Guangyu Zhang are with the College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510642, China (e-mail: guwanrong@scau.edu.cn; 20223170015@stu.scau.edu.cn; cshywang@scau.edu.cn; guangyuzhg@foxmail.com).

Bin Zhang is with the Guangdong Institute of Intelligence Science and Technology, Hengqin, Guangdong Institute of Intelligence Science and Technology, Hengqin, Zhuhai 519031, China (e-mail: zhangbin@gdiist.cn).

Jiazhou Chen and Hongmin Cai are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: csjzchen@mail.scut.edu.cn; hmcai@scut.edu.cn).

Recommended for acceptance by Y. Zhang.

Digital Object Identifier 10.1109/TETCI.2024.3502459

2471-285X © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: SOUTH CHINA UNIVERSITY OF TECHNOLOGY. Downloaded on September 29,2025 at 02:04:59 UTC from IEEE Xplore. Restrictions apply.

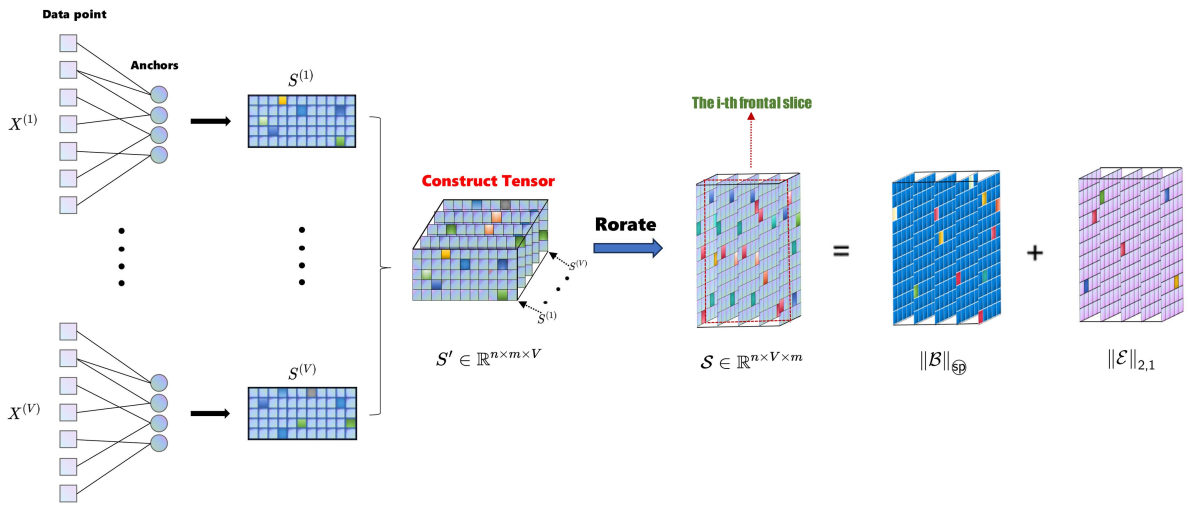


Fig. 1. The framework of ETBGL. First, in the multi-view data $X^{(v)}$, representative anchor points are selected. Then, specific bipartite graph matrices $S^{(v)} \in \mathbb{R}^{n\times m}$ are computed for each view, where n is the number of data points, m is the number of anchor points ($m \ll n$), v represents the v-th view. Next, the bipartite graph matrix is synthesized into a tensor, denoted as $S' \in \mathbb{R}^{n\times m\times V}$. Subsequently, this tensor S' undergoes a rotation operation, as illustrated in the diagram, resulting in the tensor S. Under the assumptions of low-rank and sparsity, we learn the essential tensor $\mathcal{B}$ by minimizing the tensor Schatten p-norm.

and complementarity between views. Additionally, these methods did not simultaneously consider the low-rank property and high-order structure of the fused graph, which could lead to unsatisfactory clustering results.

To address these challenges, this paper introduces a novel clustering model known as "Efficient Multi-view Clustering via Essential Tensorized Bipartite Graph Learning (ETBGL)." The fundamental framework of ETBGL is depicted in Fig. 1. This model effectively harnesses the complementary information and correlations existing between different views through the application of the Schatten p-norm to the tensorized bipartite graph. To enhance the consistency and interpretability of clustering outcomes, the model employs $\ell_{2,1}$-norm regularization on the error tensor. This regularization aids in preserving shared information within each view while mitigating the influence of noise and redundant features, consequently enhancing the overall clustering performance.

The main innovations of this paper are summarized as follows:

- We have proposed a novel method termed ETBGL, which utilizes low-rank tensor bipartite graph learning to penalize high-order information between views and capture their correlations. Moreover, ETBGL employ the $\ell_{2,1}$-norm of tensor as a sparse penalty for the error term, achieving efficient noise filtering while retaining shared information.
- An effective alternating direction method of multipliers(ADMM) scheme is designed to solve the proposed ETBGL model. Our proposed algorithm is time-economical and embraces good convergence.
- Extensive experiments on ten real-world datasets highlight that ETBGL achieves superior performances over other state-of-the-art methods.

The remainder of this paper is structured as follows. Section II introduces some notations used in this paper and briefly overviews the preliminary works for this paper. Section III introduces an original essential tensor learning model and its optimization algorithm. Section IV solves the model optimization problem. Section V provides the experimental results on ten real-world datasets and analysis of the proposed algorithm. Section VI concludes this paper.

II. Notations and Related Work

A. Notations

Throughout the paper, we use bold calligraphy letters (e.g., $\mathcal{A} \in \mathbb{R}^{n_1\times n_2\times n_3}$) for 3rd-order tensors and the bold uppercase characters ($e.g., A \in \mathbb{R}^{n_1\times n_2}$) for matrices, respectively. For a 3rd-order tensor $\mathcal{A} \in \mathbb{R}^{n_1\times n_2\times n_3}$, its k-th frontal slice is written as $\mathcal{A}(:,:,k)$. For simplicity, the $\mathcal{A}(:,:,k)$ is often written in the form of $\mathcal{A}^{(k)}$. Furthermore, the mode-1, mode-2, and mode-3 fibers of tensor $\mathcal{A} \in \mathbb{R}^{n_1\times n_2\times n_3}$ are written as $\mathcal{A}(:,j,k)$, $\mathcal{A}(i,:,k)$, and $\mathcal{A}(i,j,:)$.

In this article, the $\overline{\mathcal{A}}$ denotes the discrete Fast Fourier Transform (FFT) of $\mathcal{A}$ along the third dimension, e.g., $\overline{\mathcal{A}} = \text{fft}(\mathcal{A}, [], 3)$. Thus, $\mathcal{A} = \text{ifft}(\overline{\mathcal{A}}, [], 3)$. The $\ell_{2,1}$-norm of a tensor is defined as the sum of ℓ_2-norm of each mode-3 fiber. e.g., $\|\mathcal{A}\|_{2,1} = \sum_i^j \|\mathcal{A}(i,j,:)\|_2$.

Definition 1. (Frobenius Norm and Nuclear Norm): For a matrix $A \in \mathbb{R}^{n_1\times n_2}$, its Frobenius norm is defined as $\|A\|_F = (\sum_{i,j} |x_{i,j}|^2)^{1/2}$; its nuclear norm $\|A\|_*$ equals the sum of singular values of A.

Definition 2. (Tensor Nuclear Norm) [24]: For a 3rd-order tensor $\mathcal{A} \in \mathbb{R}^{n_1\times n_2\times n_3}$, Suppose h equals $min(n_1, n_2)$, its tensor nuclear norm is written as:

$$\begin{aligned} &\|\mathcal{A}\|_{TNN} \\ &= \sum_{i=1}^{n_3} \|\overline{\mathcal{A}}^{(i)}\|_* \\ &= \sum_{i=1}^{n_3}\sum_{j=1}^{h} \sigma_j\left(\overline{\mathcal{A}}^{(i)}\right). \end{aligned} \tag{1}$$

Authorized licensed use limited to: SOUTH CHINA UNIVERSITY OF TECHNOLOGY. Downloaded on September 29,2025 at 02:04:59 UTC from IEEE Xplore. Restrictions apply.

TABLE I
NOTATION

Notation	Explanation
A	A Matrix
$\mathcal{A}$	3rd-order tensors
$\overline{\mathcal{A}}$	The discrete Fast Fourier Transform (FFT) of $\mathcal{A}$ along the third dimension
$\|\mathcal{A}\|_{2,1}$	$\|\mathcal{A}\|_{2,1} = \sum_i^j \|\mathcal{A}(i,j,:)\|_2$
$\|A\|_\infty$	$\|A\|_\infty = \max_{ijk} \|A_{ijk}\|$
$\|\mathcal{A}\|_{TNN}$	The tensor nuclear norm of a tensor $\mathcal{A}$
$\|\mathcal{A}\|_{\circledS p}$	The tensor Schatten p-norm of a tensor $\mathcal{A}$

Definition 3. (Tensor Schatten p-norm) [26]: For a 3rd-order tensor $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, Suppose h equals $min(n_1, n_2)$, the tensor Schatten p-norm of tensor $\mathcal{A}$ is defined as:

$$\|\mathcal{A}\|_{\circledS p} = \left(\sum_{i=1}^{n_3} \left\| \overline{\mathcal{A}}^{(i)} \right\|_{\circledS p}^p \right)^{\frac{1}{p}} = \left(\sum_{i=1}^{n_3} \sum_{j=1}^{h} \sigma_j \left(\overline{\mathcal{A}}^{(i)} \right)^p \right)^{\frac{1}{p}}. \quad (2)$$

Remake 1: Based on the properties of the Schatten p-norm, it can be determined that when $p = 1$, the tensor Schatten p-norm becomes equivalent to the tensor nuclear norm: $\|\mathcal{A}\|_{TNN} = \sum_{i=1}^{n_3} \sum_{j=1}^{h} \sigma_j(\overline{\mathcal{A}}^{(i)})$. For easy of representation, we leverage matrix instead of tensor to explain the Schatten p-norm. Given matrix $A \in \mathbb{R}^{n_1 \times n_2}$ and its singular values $\sigma_1, \ldots, \sigma_h$ which are sorted in descending order. When $p > 0$, we have $\|\mathcal{A}\|_{\circ sp}^p = \sigma_1^p + \cdots + \sigma_h^p$. When $p \to 0$, we have $\lim_{p \to 0} \|\mathcal{A}\|_{\circ sp}^p = \{\sharp i : \sigma_i \neq\} = rank(A)$.

To facilitate easy reference, we have summarized the commonly used symbols in Table I.

B. Recent Advances in Multi-View Clustering Learning

Multi-view clustering methods have long garnered widespread attention. The fundamental steps can be summarized as follows: initially constructing a similarity matrix from the data, and then utilizing the properties of the graph Laplacian to learn the affinity matrix. Classic approaches such as Spectral Clustering (SPC) [15] and Co-regularization (Co-reg) [16] exemplify this process, and many contemporary clustering methods are derived from their foundations. Within these methods, a further subdivision can be made into two categories: methods based on common eigenvector matrix [26], [27] and methods based on common coefficient matrices [10], [28].

In clustering models, the utilization of low-rank representation plays a crucial role in capturing underlying structures within the data and mitigating the risk of overfitting. W. Liu et al. [17] introduced a method that captures high-order relationships between multiple-view data through tensor low-rank constraints. Q. Zhou et al. [18] employed tensor Schatten p-norm minimization and nuclear norm minimization to jointly explore low-rank structures embedded within inter-view graphs and intra-view graphs. Similarly, R. Xia et al. [19] and J. Wu et al. [27] employed a Markov chain subspace method incorporating low-rank constraints and sparse decomposition to extract consistent information from multiple views. M. BrbiÄ et al. [20] utilized both sparsity and low-rank exploration on the shared affinity matrix to uncover complementary structures, thereby enhancing consistent subspace patterns and improving clustering effectiveness. However, it's worth noting that these aforementioned approaches, while addressing inter-view complementarity and consistency, often overlooked the incorporation of high-order information inherent in multi-view data. Additionally, they may become inefficient in terms of computation and storage when confronted with large-scale scenario challenges.

To address this issue, methods based on bipartite graphs have become a focal point of research [9], [10]. For instance, H. Chen et al. [13] proposed selecting representative data points as anchors to generate bipartite graphs instead of using traditional similarity graphs. However, merely relying on kernel methods for approximation does not provide a truly coefficient representation. L. Li et al. [12] and Z. Kang et al. [10] employed adaptive learning of different view weights by fusing bipartite graph matrices. Nevertheless, their approach of simply learning anchor graphs does not sufficiently explore the effective information within views. Similarly, S. Wang et al. [28] and S. Fang et al. [30] integrated anchor selection and subspace graph construction into a unified optimization framework, simplifying the learning steps. However, they did not effectively explore the valuable information embedded across views, resulting in an insufficient extraction of information contained within each view. H. Yang et al. [25] utilized the constructed bipartite graph to form a tensor and effectively encoded complementary information among views through Schatten p-norm regularization. Q. Gao et al. [24] and W. Xia et al. [26] introduced the tensor Schatten p-norm, penalizing singular values to encourage low-rank approximation and better capture the primary structure within the data. This approach avoids potential information loss in tensor nuclear norm regularization. However, it requires balancing the model's complexity to prevent the model from becoming overly complex by introducing additional modules.

III. Efficient Multi-View Clustering Via Essential Tensorized Bipartite Graph Learning

In this section, we begin by presenting an overview of the bipartite graph-based multi-view clustering model. Subsequently, we provide a detailed exposition of our proposed ETBGL model.

A. Multi-View Clustering With Bipartite Graphs

Given the inherent advantage of bipartite graph-based methods in efficiently reducing the computational complexity of generating similarity graphs and effectively integrating multiple view information, several bipartite graph-based multi-view clustering approaches have been proposed [12], [22]. For multi-view data, $X^{(v)} \in \mathbb{R}^{n \times d(v)} (v = 1, \ldots, V)$ denotes the data matrix of the v-th view, $d^{(v)}$ is the feature dimension of v-th view. The multi-view clustering method based on bipartite graphs can be

Authorized licensed use limited to: SOUTH CHINA UNIVERSITY OF TECHNOLOGY. Downloaded on September 29,2025 at 02:04:59 UTC from IEEE Xplore. Restrictions apply.

succinctly expressed as [23]:

$$\min_{\mathrm{P^TP=I},\delta^{(v)}} \sum_{v=1}^{V} \delta^{(v)} \mathrm{Tr}(\mathrm{P^T}\widetilde{\mathrm{L}}^{(\mathrm{v})}\mathrm{P}) + \varphi\mathcal{R}(\delta),$$
$$\text{s.t.} \sum_{v=1}^{V} \delta^{(v)} = 1, \delta^{(v)} \geq 0. \tag{3}$$

where $\delta = [\delta^{(1)}, \ldots, \delta^{(V)}]$ is the weighted vector and $\delta^{(v)}$ is the weight of the v-th view; $\mathcal{R}(\cdot)$ serves as a regularizer to balance the model's fitting capability and generalization ability. φ is a trade-off parameter. $\mathrm{Tr}(\cdot)$ represents the trace operation on a matrix. $\widetilde{\mathrm{L}}^{(\mathrm{v})} = \mathrm{I} - (\mathrm{D}^{(\mathrm{v})})^{-\frac{1}{2}}\mathrm{Z}^{(\mathrm{v})}(\mathrm{D}^{(\mathrm{v})})^{-\frac{1}{2}}$ is the normalized Laplacian matrix of v-th view, where $\mathrm{Z}^{(\mathrm{v})} = \begin{bmatrix} 0 & B^{(v)} \\ B^{(v)} & 0 \end{bmatrix}$ is full adjacency matrix of the bipartite graph;the optimal solution P is obtained by the k eigenvectors of $\widetilde{\mathrm{L}}$ corresponding to the k smallest eigenvalues.

B. Tensorized Bipartite Graph Learning

Although the aforementioned bipartite graph models have gained substantial attention, there are still numerous challenges awaiting resolution. For instance, they often fail to simultaneously account for both intra-view and inter-view similarities and struggle to effectively handle more intricate high-order information inherent in multi-view clustering tasks. Addressing this, Yang et al. [25] tackled the problem by utilizing tensors to represent the high-order information between multiple views:

$$\min_{\mathrm{P^TP=I},\delta^{(v)},\mathcal{B}} \|\mathcal{B}\|_{\circledS p}^{p} + \sum_{v=1}^{v} \delta^{(v)} \mathrm{Tr}(\mathrm{P^T}\widetilde{\mathrm{L}}^{(\mathrm{v})}\mathrm{P}),$$
$$\text{s.t.} \sum_{v=1}^{V} \delta^{(v)} = 1, \delta^{(v)} \geq 0, \mathrm{B}^{(\mathrm{v})}1 = 1, \mathrm{B}^{(\mathrm{v})} \geq 0. \tag{4}$$

where $\mathcal{B}(:, v, :) = B^{(v)}$ represents a $n \times m$ tensor graph. In this context, F stands for the cluster indicator matrix, and the optimal solution for F comprises the eigenvectors corresponding to the smallest eigenvalues of $\widetilde{\mathrm{L}}^{(\mathrm{v})}$.

The MSC-BG model synthesizes tensors by combining data matrices and applies Schatten p-norm constraint to explore higher-order information between views. However, some view information may contain redundant details, such as noise or errors caused by inaccurate data. These factors can affect the accuracy of the results. Although the model considers the similarity between views, it ignores the influence of error information in views.

C. ETBGL Model

To solve the above problems, we propose ETBGL model. For multi-view data, denote by $X^{(v)} \in \mathbb{R}^{n \times d^{(v)}} (v = 1, \ldots, V)$ the data matrix of the v-th view, V is the number of views, $d^{(v)}$ is the feature dimension of v-th view. The final objective function of our proposed ETBGL method can be reformulated as follows:

$$\min_{\mathcal{B},\mathcal{E}} \|\mathcal{B}\|_{\circledS p} + \lambda \|\mathcal{E}\|_{2,1}$$
$$s.t. \quad \mathcal{S} = \mathcal{B} + \mathcal{E}. \tag{5}$$

where $\mathcal{S}$ denotes the bipartite graph tensor, $\mathcal{B}$ represents the essential tensor learned under the low-rank and sparsity assumptions.In our approach, when synthesizing the tensor from the data matrix, we perform tensor decomposition to produce the essential tensor $\mathcal{B}$ and the error tensor $\mathcal{E}$. The final graph is then obtained through the decomposition of the essential tensor. To mitigate the influence of noise and outliers in the samples on bipartite graph tensor learning, we introduce an $\ell_{2,1}$-norm to the error term [27]. The tensor $\ell_{2,1}$-norm allows us to impose an ℓ_2-norm constraint on the error matrix of each view within the tensor. This constraint ensures that the error matrix of each view remains as small as possible. Given that noisy samples exhibit sparsity, employing the $\ell_{2,1}$-norm and low-rank representation are particularly suitable for alleviating noise and outliers to learn a "clean" representation [37]. By addressing the error terms within each view, this model can simultaneously consider both inter-view and intra-view sample similarities, thereby better exploring the embedded information within each view.

In this paper, the anchor method adopts the anchor selection technique from reference [26], namely VDA (Variance-based De-correlation Anchor selection), to select anchors for the model. This method begins by computing the variance θ for each sample's d-dimensional features. Then, it normalizes this variance by dividing it by the maximum variance in the score vector. It selects the data point with the highest score as the initial anchor point and iteratively chooses data points with the highest scores as anchors through a cyclic strategy.

IV. Optimization

We adopt the alternating direction method of multipliers (ADMM) to solve (5). ADMM decomposes the original problem into multiple subproblems and solves them alternately by optimizing the variables. This decomposition property makes ADMM particularly effective for problems with structured constraints or distributed optimization. The augmented Lagrangian function can be formulated as follows:

$$\begin{aligned}\mathcal{L}(\mathcal{B},\mathcal{E}) =& \|\mathcal{B}\|_{\circledS p}^{p} + \lambda \|\mathcal{E}\|_{2,1} \\ &+ < \mathcal{Y}, \mathcal{S} - \mathcal{B} - \mathcal{E} > + \frac{\rho}{2} \|\mathcal{S} - \mathcal{B} - \mathcal{E}\|_F^2 \\ =& \|\mathcal{B}\|_{\circledS p}^{p} + \lambda \|\mathcal{E}\|_{2,1} + \frac{\rho}{2} \left\| \mathcal{S} - \mathcal{B} - \mathcal{E} + \frac{\mathcal{Y}}{\rho} \right\|_F^2,\end{aligned} \tag{6}$$

where $\rho > 0$ is a parameter that controls the strength of the penalty term and $\mathcal{Y}$ is the Lagrange multiplier. By continuously updating $\mathcal{Y}$, the ADMM algorithm progressively approaches the optimal solution of the original optimization problem.

1) $\mathcal{E}$ sub-problem, solving $\mathcal{E}$ with fixed $\mathcal{B}$

$$\mathcal{E}^{k+1} = arg\min_{\mathcal{E}} \lambda \|\mathcal{E}\|_{2,1} + \frac{\rho^k}{2} \left\| \mathcal{E} - \left(\mathcal{S} - \mathcal{B}^k + \frac{\mathcal{Y}^k}{\rho^k} \right) \right\|_F^2. \tag{7}$$

Authorized licensed use limited to: SOUTH CHINA UNIVERSITY OF TECHNOLOGY. Downloaded on September 29,2025 at 02:04:59 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1: Algorithm of Essential Tensorized Bipartite Graph Learning (ETBGL).

Input: Data matrices $\{X^{(v)}\}_{v=1}^{V} \in \mathbb{R}^{n\times d^{(v)}}$, anchors number m, number of clusters c, regularization parameters λ.
Output: $\mathcal{B}^{k+1}$, $\mathcal{E}^{k+1}$, and c Clusters.

```
Organize anchor matrices {S^(v)}_{v=1}^V into tensorial data S.
Initialize k = 0, B^0 = E^0 = Y^0 = 0, ρ^0 = 10^-3, ρ^max = 10^10, η = 1.9, and γ = 10^-6.
while not converged do
    Fix E^k. Updata B^{k+1} by Eq. (13)
    Fix B^{k+1}. Updata E^{k+1} by Eq. (9)
    Updata Y^{k+1} by Eq. (14)
    ρ^{k+1} = min(ηρ^k, ρ^max).
    Check the convergence conditions:
    ||B^{k+1} − B^k||_∞ ≤ γ,
    ||E^{k+1} − E^k||_∞ ≤ γ,
    ||S − B^{k+1} − E^{k+1}||_∞ ≤ γ.
    k = k + 1.
end while
Compute B* = (1/V) Σ_{v=1}^V |B(:, v, :)|.
Obtain H by performing SVD on B*.
Perform k-means on H to achieve the final partition.
```

As the $\ell_{2,1}$-norm of the tensor $\mathcal{E}$ is defined as the sum of ℓ_2-norm of each mode-3 fiber. To conveniently calculate the ℓ_2-norm of each mode-3 fiber and sum them up, we perform an operation by matricizing each tensor $\mathcal{E}^{(k+1)}$ along the 3rd mode. So we have $\|E_{(3)}^{k+1}\|_{2,1} = \|\mathcal{E}^{k+1}\|_{2,1}$. Then the formula (7) can be converted to:

$$E_{(3)}^{k+1} = arg\min_{E_{(3)}} \lambda \left\|E_{(3)}^{k+1}\right\|_{2,1} + \frac{\rho^k}{2}\left\|E_{(3)} - \left(S_{(3)} - B_{(3)}^k + \frac{Y_{(3)}^k}{\rho^k}\right)\right\|_F^2. \quad (8)$$

Let $Q = S_{(3)} - B_{(3)}^k + \frac{Y_{(3)}^k}{\rho^k}$. According to [27], the problem in (8) has the following close-form solution:

$$\mathrm{E}_{(3):,i}^{k+1} = \begin{cases} \frac{\|Q_{:,i}\|_2 - \frac{\lambda}{\rho^k}}{\|Q_{:,i}\|_2} Q_{:,i}, & if\ \|Q_{:,i}\|_2 > \frac{\lambda}{\rho^k} \\ 0, & otherwise, \end{cases} \quad (9)$$

where $Q_{:i}$ represents the i-th column of the matix Q. This allows us to easily compute the 2-norm for each column and obtain the tensor $\mathcal{E}$.

2) $\mathcal{B}$ sub-problem, solving $\mathcal{B}$ with fixed $\mathcal{E}$

$$\mathcal{B}^{k+1} = arg\min_{\mathcal{B}} \|\mathcal{B}\|_{\circledS p}^p + \frac{\rho^k}{2}\left\|\mathcal{B} - \left(\mathcal{S} - \mathcal{E}^k + \frac{\mathcal{Y}^k}{\rho^k}\right)\right\|_F^2. \quad (10)$$

The problem in (10) can be computed by [24]. As we know, Theorem 3 [24]: Suppose $\mathcal{D} \in \mathcal{R}^{n_1\times n_2\times n_3}$, let $\mathcal{D} = \mathcal{U} * \mathcal{S} * \mathcal{V}^T$. Given the model:

$$arg\min_{\mathcal{X}} \frac{1}{2}\|\mathcal{X} - \mathcal{D}\|_F^2 + \tau\|\mathcal{X}\|_{\circledS p}^p. \quad (11)$$

Then, the minimizer solution of (11) is:

$$\mathcal{X}^* = \Gamma_{\tau * n_3}(\mathcal{D}) = \mathcal{U} * \mathrm{ifft}(\mathrm{P}_{\tau * \mathrm{n}_3}(\overline{\mathcal{D}})) * \mathcal{V}^T, \quad (12)$$

where $P_{\tau * n_3}(\overline{\mathcal{D}})$ is a tensor, whose v-th frontal slice $P_{\tau * n_3}(\overline{\mathcal{D}}^{(v)})$ can be solved via using the generalized shrinkage-thresholding (**GST**) algorithm.

Analysis of the above **Theorem 3** [24] can directly get the optimal solution of the model (10):

$$\mathcal{B}^{k+1} = \Gamma_{\frac{1}{\rho^k}}\left(\mathcal{S} - \mathcal{E}^k + \frac{\mathcal{Y}^k}{\rho^k}\right). \quad (13)$$

3) Updata Lagrange multiplie

$$\mathcal{Y}^{k+1} = \mathcal{Y}^k + \rho^k(\mathcal{S} - \mathcal{B}^{k+1} - \mathcal{E}^{k+1}). \quad (14)$$

$$\rho^{k+1} = \min(\eta\rho^k, \rho_{max}). \quad (15)$$

The complete procedure for solving $\mathcal{B}$ and the steps of our method are summarized in Algorithm 1, and the stopping condition is set by $\max\ (\|\mathcal{B}^{k+1} - \mathcal{B}^k\|_\infty, \|\mathcal{E}^{k+1} - \mathcal{E}^k\|_\infty, \|\mathcal{S} - \mathcal{B}^{k+1} - \mathcal{E}^{k+1}\|_\infty) \le \gamma$.

A. Complexity Analyses

The Computational complexity of our method consists of two main components: **1)** Constructing the graph $\{B^{(v)}\}_{v=1}^V$, which involves constructing the graph B and spends $O(Vnmd + Vnm\log(m))$, where $d = \sum_{v=1}^V d^{(v)}$, and V, m, and n represent the number of views, anchor points, and samples, respectively, with $m \ll n$. **2)** Optimizing through iterative solving of equation (5). In this part, the focus is on two variables, $\mathcal{B}$ and $\mathcal{E}$. Each iteration requires $O(Vnm)$ time to compute an approximate solution for $\mathcal{E}$. Given K as the number of iterations, the overall computation for approximating $\mathcal{E}$ requires $O(KVnm)$ time. Regarding variable $\mathcal{B}$, solving its sub-problem involves calculating the FFT and inverse FFT of an $n \times V \times m$ tensor along the third dimension and performing n SVDs of $m \times V$ matrices in the Fourier domain. Both of these operations have complexities of $O(Vnm\log(Vm))$ and $O(V^2mn)$, respectively. Hence, the total computational cost of updating $\mathcal{B}$ with K iterations is $O(K(Vnm\log(Vm) + V^2mn))$. Overall, the primary computational complexity of ETBGL is $O(Vnmd + KVnm\log(VM))$. Due to the condition where $m \ll n$, our method exhibits a linear relationship in terms of time complexity with respect to n .

V. Experiments

In this section, we tested the ETBGL model and other comparative models through a large amount of data to verify their effectiveness. Our experimental platform is the Windows 11 platform, with an i5-12400F processor and a MATLAB version of R2022a.

Authorized licensed use limited to: SOUTH CHINA UNIVERSITY OF TECHNOLOGY. Downloaded on September 29,2025 at 02:04:59 UTC from IEEE Xplore. Restrictions apply.

TABLE II
EXPERIMENTAL DATASETS

Dataset	Samples	Objective	View	Clusters
BBC-Sport	544	Text	2	5
Prokaryotic	551	Biology	3	4
Pascals	1000	Images/Text	2	20
COIL20	1440	Images	3	20
HW1256	2000	Images	4	10
UCI-Digits	2000	Images	3	10
Mnist4	4000	Images	3	4
Scene15	4485	Images	3	15
BioData_2K3K	9190	Biology	2	22
Hdigit	10000	Images	2	10

A. Datasets

We conducted experiments on ten multi-view datasets. In Table II, we summarized the basic information of these datasets.

- **BBC-Sport**:[1] This dataset is divided into two views, which view has 544 documents embracing five main sports news topics sourced from the BBC-Sport website.
- **Prokaryotic** [37]: This dataset consists of 551 prokaryotic species, which are distributed across four imbalanced clusters. These species are characterized using two distinct genomic representations: the first being the composition of 3 proteomes, the second consisting of a repertoire of 393 genes. Additionally, textual data describing prokaryotic species is included, represented as a bag-of-words representation comprising 438 features.
- **Pascals** [36]: This dataset is an image/text dataset selected from 2008 PASCAL development kit. In this dataset, each image is described by 5 sentences. It has 1000 image/text pairs belonging to 20 categories totally.
- **COIL20**:[2] This dataset is the abbreviation of the Columbia object image library dataset, which contains 1,440 images of 20 object categories. Each category contains 72 images and all images are normalized to size 32Ã32.
- **HW1256** [34]: This dataset includes two thousand digital image samples representing numbers 0-9. The Fourier coefficients of 76 dimensional feature shapes, 216 dimensional contour related features, and 64 dimensional Karhunen-love coefficient features were selected to construct the three views datasets.
- **UCI-Digits** [31]: This dataset consists of 2000 handwritten digit images, divided into 3 views, each corresponding to 10 classes.
- **Mnist4** [33]: This dataset comprises 4000 handwritten digit images, categorized into three views, each representing four classes of handwritten digits.
- **Scene15** [35]: The dataset comprises 4,485 images encompassing 15 natural scene categories, categorized into indoor and outdoor environments. These include images from industrial settings, shops, bedrooms, kitchens, and more.
- **DioData_2K3K**:[3] This dataset is a multi-omics dataset consisting of 2000-dimensional RNA-seq and 3000-dimensional ATAC expression profiles of mouse cortex from the Chen-2019 database.
- **Hdigit** [32]: The dataset consists of two views sourced from MNIST and USPS handwritten digit datasets. In total, there are 10,000 handwritten digit images, divided into ten categories.

B. Compared Methods

We compared five state-of-the-art multi-view clustering methods with our proposed ETBGL algorithm as stated below:

1) **ETLMSC**[2019,T-IP] [27]: This work proposes a novel essential tensor learning method, constructing a multi-view tensor with Markov chain-based transition probability matrices and seeking the low-rank representation of the essential tensor by applying the tensor-singular value decomposition (t-SVD) based tensor nuclear norm.
2) **LMVSC**[2020,AAAI-20] [10]: This work introduces a large-scale multi-view subspace clustering (LMVSC) algorithm with linear order complexity. LMVSC learns a smaller graph between the raw samples and the anchor points for each view and designs a novel way to integrate those graphs to implement clustering in large-scale scenarios.
3) **FPMVS-CAG**[2021,T-IP] [28]: This work presents a fast parameter-free multiview subspace clustering method with consensus anchor guidance (FPMVS-CAG), which jointly optimizes anchor selection and subspace graph construction in a unified framework.
4) **TBGL**[2022,TPAMI] [26]: This work provides a tensorized bipartite graph learning for multi-view clustering (TBGL), which simultaneously exploits inter-view and intra-view similarities with tensor Schatten p-norm and $\ell_{1,2}$-norm penalty.
5) **UDBGL**[2023,TNNLS] [30]: This work presents an efficient MVC approach via unified and discrete bipartite graph learning (UDBGL), which integrates view-specific and view-consensus bipartite graph learning with discrete cluster structure learning into a unified objective function and learns view-specific bipartite graphs from multiple perspectives.

C. Evaluation Metrics

To comprehensively evaluate clustering performance, we utilize all six commonly used metrics, which include Accuracy (ACC) [12], Normalized Mutual Information (NMI) [38], F-score [39], Precision [10], Recall [27], and Purity [23]. These six metrics capture different aspects of clustering tasks. It's important to note that, for all these metrics, higher values indicate better performance.

D. Performance Comparison

Tables III–VII present the comparative results of all methods on ten datasets, where all results are measured as the maximum

[1][Online]. Avilable: http://mlg.ucd.ie/datasets/segment.html
[2][Online]. Avilable: http://www.cs.columbia.edu/CAVE/software/softlib/
[3][Online]. Avilable: https://scglue.readthedocs.io/zh_CN/latest/data.html

Authorized licensed use limited to: SOUTH CHINA UNIVERSITY OF TECHNOLOGY. Downloaded on September 29,2025 at 02:04:59 UTC from IEEE Xplore. Restrictions apply.

TABLE III
EXPERIMENT RESULTS ON THE COIL20 AND HW1256 DATASETS

	COIL20						HW1256					
Methods	ACC	NMI	Purity	F-score	Precision	Recall	ACC	NMI	Purity	F-score	Precision	Recall
ETLMSC	0.8472	0.8919	0.8625	0.7998	0.7815	0.8189	0.8440	0.8711	0.8615	0.8136	0.7798	0.8795
LMVSC	0.7174	0.7990	0.7243	0.6535	0.6352	0.6880	0.7355	0.6912	0.7385	0.6261	0.6142	0.7044
FPMVS-CAG	0.5917	0.7898	0.6424	0.5890	0.5013	0.7139	0.7425	0.7490	0.7425	0.6752	0.6353	0.7653
TBGL	0.6146	0.7581	0.6479	0.5231	0.4742	0.6100	0.9215	0.9000	0.9215	0.8732	0.8665	0.9142
UDBGL	0.6340	0.7944	0.6868	0.6204	0.5671	0.6849	0.8180	0.7771	0.8180	0.7472	0.7062	0.7932
ETBGL	**0.9535**	**0.9459**	**0.9535**	**0.9136**	**0.9107**	**0.9165**	**0.9990**	**0.9973**	**0.9990**	**0.9980**	**0.9980**	**0.9980**

For ETBGL,we set anchor_rate=0.016,λ = 0.05 and anchor_rate=0.012,λ = 0.1 for these two datasets.

TABLE IV
EXPERIMENT RESULTS ON THE UCI-DIGITS AND MNIST4 DATASETS

	UCI-DIGITS						Mnist4					
Methods	ACC	NMI	Purity	F-score	Precision	Recall	ACC	NMI	Purity	F-score	Precision	Recall
ETLMSC	0.8945	0.8580	0.8945	0.8292	0.8271	0.8435	0.9370	0.8535	0.9370	0.8860	0.8831	**0.8889**
LMVSC	0.8510	0.7420	0.8510	0.7238	0.7099	0.6926	0.7962	0.5844	0.7962	0.6826	0.6683	0.6976
FPMVS-CAG	0.7310	0.7467	0.7350	0.6891	0.6431	0.7713	0.8848	0.7145	0.8848	0.7993	0.7941	0.8085
TBGL	0.8340	0.8368	0.8340	0.7834	0.7747	0.8267	0.9197	0.8073	0.9197	0.8566	0.8523	0.8609
UDBGL	0.8265	0.8053	0.8275	0.7634	0.7177	0.8196	0.8758	0.7044	0.8758	0.7848	0.7762	0.8049
ETBGL	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**0.9605**	**0.8940**	**0.9605**	**0.9255**	**0.9234**	0.3923

For ETBGL,we set anchor_rate=0.011,λ = 0.2 and anchor_rate=0.005,λ = 0.09 for these two datasets.

TABLE V
EXPERIMENTS RESULTS ON THE BIODATA_2K3K AND PASCALS DATASETS

	BioData_2K3K						Pascals					
Methods	ACC	NMI	Purity	F-score	Precision	Recall	ACC	NMI	Purity	F-score	Precision	Recall
ETLMSC	0.3090	0.4460	0.5303	0.3209	0.4803	0.2484	0.8860	**0.9132**	0.9020	0.8476	0.8161	**0.8816**
LMVSC	0.1402	0.0194	0.2458	0.1436	0.1203	0.1830	0.4030	0.4204	0.4170	0.2597	0.2514	0.2727
FPMVS-CAG	0.3726	0.2975	0.4446	0.2722	0.2647	0.3006	0.3760	0.4384	0.3970	0.2826	0.2484	0.3319
TBGL	0.2655	0.0037	0.2667	0.2364	0.1341	**0.9966**	0.8130	0.8293	0.8330	0.7399	0.7186	0.7726
UDBGL	0.2819	0.2598	0.4135	0.2030	0.1857	0.2237	0.3790	0.4755	0.4030	0.2913	0.2400	0.3704
ETBGL	**0.4076**	**0.5091**	**0.6235**	**0.4579**	**0.5585**	0.3948	**0.9140**	0.9025	**0.9140**	**0.8512**	**0.8489**	0.8535

For ETBGL,we set anchor_rate=0.005,λ = 0.09 and anchor_rate=0.027,λ = 0.01 for these two datasets.

TABLE VI
EXPERIMENT RESULTS ON THE HDIGIT AND OUTSENEC DATASETS

	Hdigit						Scene15					
Methods	ACC	NMI	Purity	F-score	Precision	Recall	ACC	NMI	Purity	F-score	Precision	Recall
ETLMSC	0.9880	0.9687	0.9880	0.9762	0.9761	0.9763	0.7512	0.7886	0.6824	0.6882	0.6824	0.6942
LMVSC	0.6540	0.5491	0.6596	0.5258	0.5143	0.5475	0.5877	0.5357	0.6238	0.4530	0.4493	0.4617
FPMVS-CAG	0.9104	0.8163	0.9104	0.8334	0.8303	0.8602	0.5302	0.5718	0.5485	0.4271	0.3580	0.5510
TBGL	0.8744	0.7668	0.8744	0.7670	0.7576	0.7766	0.5211	0.4996	0.5563	0.3797	0.3671	0.3942
UDBGL	0.8778	0.7825	0.8778	0.7833	0.7800	0.8175	0.5958	0.5703	0.6167	0.4449	0.4440	0.4459
ETBGL	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**0.9795**	**0.9652**	**0.9681**	**0.9677**	**0.9684**	**0.9657**

For ETBGL,we set anchor_rate=0.014,λ = 0.08 and anchor_rate=0.004,λ = 0.1 for these two datasets.

TABLE VII
EXPERIMENTS RESULTS ON THE BBC-SPORT AND PROKARYOTIC DATASETS

	BBC-Sport						Prokaryotic					
Methods	ACC	NMI	Purity	F-score	Precision	Recall	ACC	NMI	Purity	F-score	Precision	Recall
ETLMSC	**0.9908**	**0.9737**	**0.9908**	**0.9873**	**0.9912**	**0.9835**	0.5735	0.0193	0.5735	0.5677	0.3970	**0.9961**
LMVSC	0.7794	0.6485	0.7977	0.6444	0.6746	0.6376	**0.8349**	**0.5071**	**0.8367**	**0.7745**	0.7481	0.8385
FPMVS-CAG	0.4062	0.1058	0.4816	0.3131	0.3232	0.3273	0.4301	0.1525	0.6025	0.3618	0.4656	0.2959
TBGL	0.9779	0.9258	0.9779	0.9566	0.9632	0.9585	0.7441	0.4020	0.8167	0.6625	0.6937	0.6935
UDBGL	0.4963	0.2309	0.4982	0.4856	0.3628	0.7474	0.4537	0.2195	0.6279	0.4750	0.5189	0.4379
ETBGL	0.9890	0.9612	0.9890	0.9824	0.9841	0.9807	0.7822	0.4849	0.8330	0.6840	**0.7700**	0.6153

For ETBGL,we set anchor_rate=0.052,λ = 0.09 and anchor_rate=0.1,λ = 0.13 for these two datasets.

Authorized licensed use limited to: SOUTH CHINA UNIVERSITY OF TECHNOLOGY. Downloaded on September 29,2025 at 02:04:59 UTC from IEEE Xplore. Restrictions apply.

TABLE VIII
THE WILCOXON SIGNED RANK TEST AT THE 0.01 CONFIDENCE INTERVAL LEVEL, WHERE P-VALUES ARE PROBABILITY VALUES, EVALUATES THE PROBABILITY OF THE OCCURRENCE OF THE NULL HYPOTHESIS

	P-value					
Method	ACC	NMI	Purity	F-score	Precision	Recall
ETLMSC	0.0039	0.0098	0.0039	0.0039	0.0039	0.6953
LMVSC	0.0039	0.0039	0.0039	0.0039	0.0020	0.0371
FPMVS_CAG	0.0020	0.0020	0.0020	0.0020	0.0020	0.0488
TBGL	0.0020	0.0020	0.0020	0.0020	0.0020	0.4922
UDBGL	0.0020	0.0020	0.0020	0.0020	0.0020	0.0488

value of 20 runs. The best and second-best results among the compared methods are indicated in bold and underlined, respectively. Grid search was employed to optimize the parameters for all compared methods, yielding the best results.

On all datasets, our proposed ETBGL model achieved the optimal or suboptimal results across various evaluation metrics. From Tables III–VI, one may find that ETBGL yields the competitive performance on eight benchmark datasets. ETBGL consistently outperforms all the competitors on the COIL20, HW1256, Hdigit, Scene15, UCI-Digits, Mnist4, BioData_2K3K, and Pascals datasets. Taking experiments on COIL20 as an example, our results showed improvements of 10.63%, 5.40%, 9.10%, 11.38%, 12.92%, and 9.76% over the second-best results in various metrics. ETBGL model performed exceptionally well on the large-scale biological dataset BioData_2K3K and image dataset Hdigit, showcasing its capability to handle massive datasets. This confirms that the ETBGL model is versatile and suitable for large-scale data processing tasks. Meanwhile, ETBGL achieves the second-best performance on the BBC-Sport and Prokaryotic datasets, as illustrated by Table VII. One can observe that the result of ETBGL is very close to the result of the optimal ETLMSC on the BBC-Sport. By checking the results, we can easily see that ETBGL achieves an increasing ratios of 35.21%, 20.87%, and 3.81% in terms of ACC when comparing with FPMVS-CAG, ETLMSC, and TBGL on the Prokaryotic dataset, respectively. Additionally, in Table VIII, we report the results of the Wilcoxon signed-rank test conducted at a confidence level of 0.01, comparing the ETBGL method with other methods across different evaluation metrics. The results show that, except for the Recall metric, there are significant differences between the results of the other methods and the ETBGL method in all other metrics. This further demonstrates the superiority of our approach.

The tensor-based methods, namely ETBGL, TBGL, and ETLMSC, generally outperform the multi-view subspace clustering methods, FPMVS-CAG, UDBGL, and LMVSC. This is because tensor-based methods effectively leverage the embedded information and spatial structure between views, resulting in better performance when handling different modalities of data.

ETBGL exhibits outstanding performance compared to TBGL, which is also a tensor bipartite graph-based method. The primary reason for this might lie in the different ways these two models handle the error matrix. TBGL imposes an ℓ_1-norm constraint on the error matrix to reduce unnecessary noise and redundant information. However, the $\ell_{2,1}$-norm simultaneously decreases errors within the same sample, thus retaining more feature information. In contrast, the ℓ_1-norm can shrink many samples' feature values to zero, thereby neglecting the correlations between samples. In other words, the $\ell_{2,1}$-norm takes into account the inter-sample correlations, whereas the ℓ_1-norm regularizes all samples independently without considering their inter-relationships.

Furthermore, the $\ell_{2,1}$-norm is applied to the sum of the ℓ_2-norm of each view in the tensor. If there is an outlier in one column, it will only have a significant impact on that column without affecting other columns significantly. In contrast, the ℓ_1-norm of the tensor may be more affected by outliers. The solution obtained after $\ell_{2,1}$ -norm regularization is sparser, and the features selected by different samples are more likely to be distinct. Hence, it often leads to better clustering performance.

E. Representation Visualzation

To be more intuitive, Fig. 2 depicts line plots for ACC values and bar charts for NMI results of our model and comparative models on ten benchmark datasets. As shown in Fig. 2(a), it is clear that the ETBGL (indicated by the red line) significantly outperforms other methods, with the best performance on the overwhelming majority of datasets. Similarly, the results presented in Fig. 2(b) underscore the superiority of our approach over other methods, achieving the optimal and second-best results on ten datasets.

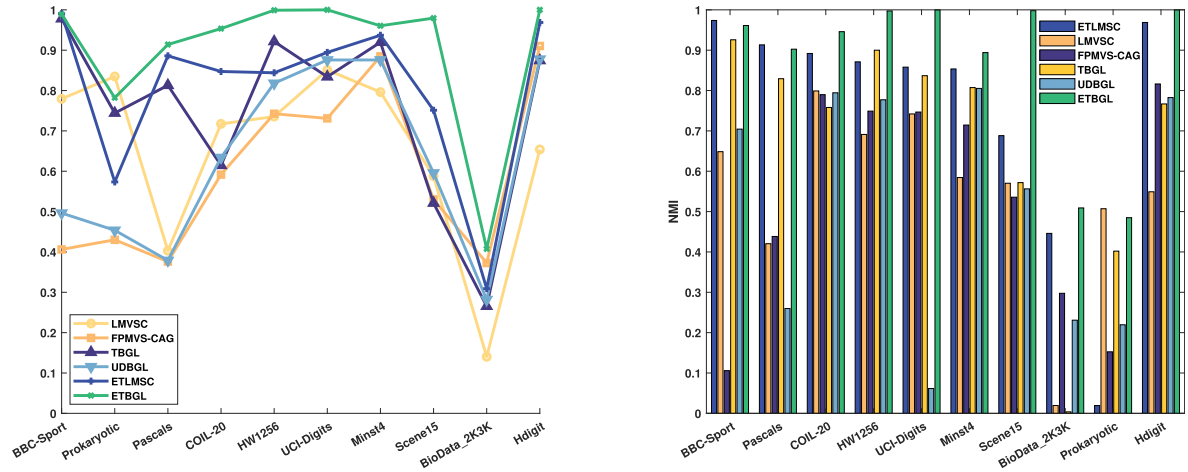
Figs. 3 and 4 present the sample spatial distribution results for the top four optimal methods on the UCI-Digits and HW1256 datasets based on the t-distributed stochastic neighborhood embedding (t-SNE) [10], respectively. t-SNE serve as a visualization tool to depict the distribution of samples in a high-dimensional space, revealing patterns of data clustering and separation. The clustering performance of all subgraphs coincided exactly with the results in Table IV. As we can observe, the ETBGL achieved the optimal separation performance.

F. Parameter Analysis

In this paper, our method primarily tunes two parameters based on grid search: anchor rate (representing the proportion of anchor points to the total number of data points) and the balancing parameter λ (pertaining to error matrix parameters). Taking UCI-Digits, HW1256, COIL20, Mnist4, BBC-Sport, and Pascals datasets as examples, we empirically analyze the effect of the anchor rate and λ for clustering on six distinct datasets in Fig. 5.

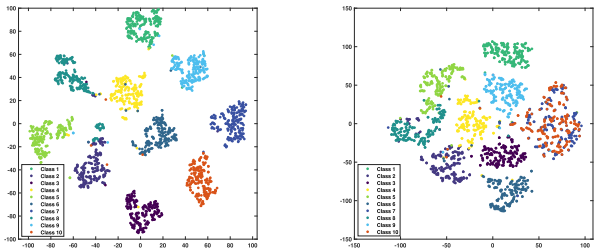
When selecting the anchor rate, we generally take into account the number of samples. In practice, having 20 to 40 anchor points often leads to excellent performance in bipartite graph clustering. Focusing on the UCI-Digits and HW1256 datasets, both of which contain 2000 samples, it is apparent from the graph that the optimal clustering performance is achieved when the anchor rate falls approximately within the range of 0.011 to 0.012. For smaller datasets like BBC-Sport or Pascals, it is recommended to keep the number of anchor points to around 20. As illustrated in Fig. 5(e) and (f) , optimal performance is

Authorized licensed use limited to: SOUTH CHINA UNIVERSITY OF TECHNOLOGY. Downloaded on September 29,2025 at 02:04:59 UTC from IEEE Xplore. Restrictions apply.



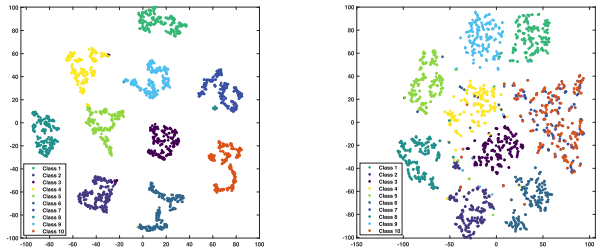
(a) (b)

Fig. 2. Clustering performance of each method on different datasets.



(a) ETBGL (b) TBGL (c) ETLMSC (d) LMVSC

Fig. 3. Visualization of clustering results using t-SNE for four methods on the UCI-Digits dataset.



(a) ETBGL (b) TBGL (c) ETLMSC (d) LMVSC

Fig. 4. Visualization of clustering results using t-SNE for four methods on the HW1256 dataset.

typically attained when the anchor rate is around 0.05 and 0.027, respectively.

The parameter λ plays a significant role in the model. When λ is small, more error terms are included in the error matrix, which can lead to increased noise and the inclusion of unnecessary information in the results. Conversely, when λ is large, more error terms are compressed to zero, resulting in sparser results with reduced noise and less unnecessary information. However, it's crucial to acknowledge that there's a trade-off, as a large λ may also result in the loss of some important information. For the ETBGL model, with a fixed anchor rate, optimal performance is observed when λ falls within the range of 0.1 to 0.12.

For the parameters ρ and η of ADMM, we directly adopt the suggestion of [27] and fix them as 10^{-3} and 1.9, respectively. These two parameters mainly influence the number of iteration for convergence.

G. Ablation Study

To further investigate the ETBGL model, ablation experiments were conducted to study the impact of the error tensor under the sparse assumption on the model. Specifically, when $\lambda = 0$, the error term is omitted, and the model directly performs t-SVD decomposition on the original tensor $\mathcal{S}$. The final clustering results are obtained by performing k-means clustering on the

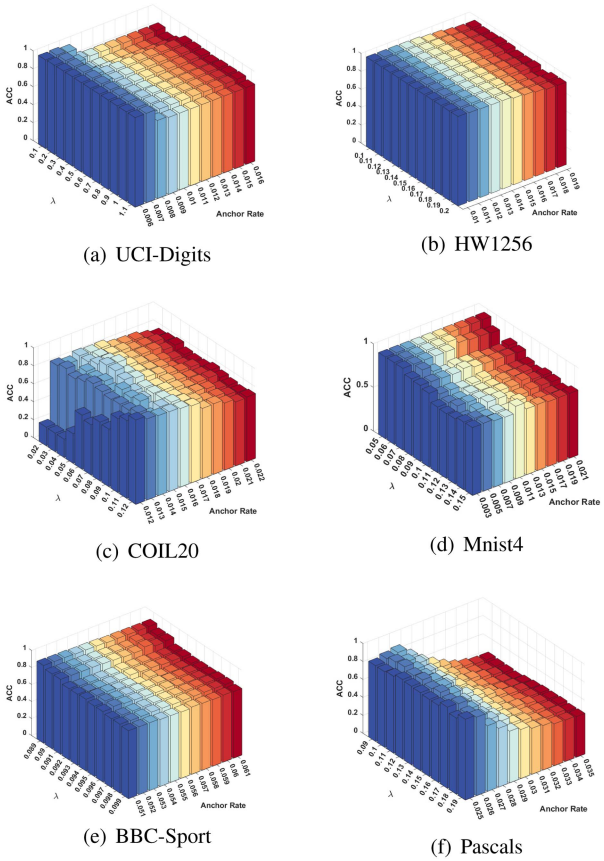
Authorized licensed use limited to: SOUTH CHINA UNIVERSITY OF TECHNOLOGY. Downloaded on September 29,2025 at 02:04:59 UTC from IEEE Xplore. Restrictions apply.

TABLE IX
THIS TABLE SUMMARIZES THE RUNTIME (IN MILLISECONDS) FOR EACH METHOD ON DIFFERENT DATASETS, WITH **"OT"** INDICATING CASES WHERE THE RUNTIME EXCEEDED **TWENTY** MINUTES

	Times(ms)									
Method	BBC-Sport	Prokaryotic	Pascals	COIL20	HW1256	UCI-Digits	Mnist4	Scene15	BioData_2K3K	Hdigit
ETLMSC	505.18	905.89	1895.46	6617.84	14687.91	12160.80	50392.98	55000.74	206437.33	220568.54
LMVSC	313.17	278.09	819.26	1096.15	1163.91	1602.33	3362.81	1655.26	19383.67	5195.70
FPMVS-CAG	5828.17	2308.57	20051.28	18787.00	5163.56	6352.65	9096.14	10971.31	88168.64	46253.07
TBGL	8004.39	111479.64	30848.19	333256.14	152125.96	81021.84	528347.09	790051.67	OT	OT
UDBGL	1534.31	1631.45	2827.34	5818.85	5297.39	5757.58	61496.46	21665.65	6655.58	21665.55
ETBGL	284.70	274.46	700.93	950.92	348.47	514.33	803.82	2165.65	14765.9	2279.16

TABLE X
ABLATION STUDY: COMPARISON RESULTS OF ETBGL AND ITS VARIANTS IN TERMS OF NMI

Variants	BBC-Sport	Prokaryotic	Pascals	COIL20	HW1256	UCI-Digits	Mnist4	Scene15	BioData_2K3K	Hdigit	AVERAGE
ETBGL	**0.9612**	**0.4849**	**0.9025**	**0.9459**	**0.9973**	**1.000**	**0.8940**	**0.9652**	**0.5091**	**1.000**	**0.8659**
ETBGL-E	0.4180	0.1540	0.7600	0.5680	0.9173	0.7875	0.5098	0.8620	0.2000	0.7664	0.5943



(a) UCI-Digits (b) HW1256 (c) COIL20 (d) Mnist4 (e) BBC-Sport (f) Pascals

Fig. 5. The impact of Hyperparameters λ and Anchor Rate on clustering performance in the UCI-Digits, HW1256, COIL20, Mnist4, BBC-Sport, and Pascals datasets.

resulting bipartite graph. This test is labeled as ETBGL-E. This test was conducted using the same parameters and experimental environment as the ETBGL model on all datasets. Table X shows the NMI values under all conditions.

As observed from Table X, the presence or absence of the error term has a significant impact on the experimental results across various datasets. For example, in the BBC-Sport, Prokaryotic, COIL20, and BioData_2K3K datasets, the influence of the error term is substantial. Additionally, calculating the NMI values across ten datasets, it can be seen that the improvement under the sparse assumption is 45.70% compared to when it is not applied. This indicates that the presence of the error term is crucial and necessary when dealing with complex and redundant datasets.

H. Analysis of the p Value

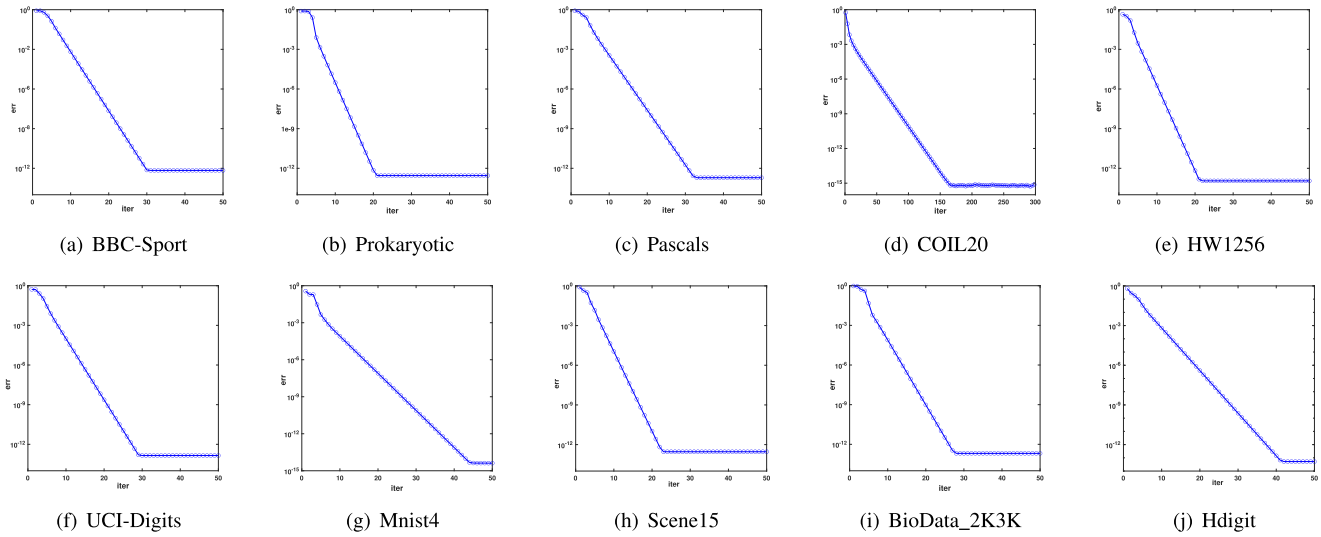
The advantage of Schatten p-norm lies in its flexibility, allowing us to choose an appropriate p value based on the specific requirements of the real problem. In Fig. 7, we analyze the impact of different p values on clustering performance in the Schatten p-norm across four datasets. In the figure, p values range from 0.1 to 1in intervals of 0.1, and each p value corresponds to three clustering evaluation metrics: ACC, NMI, and Purity.

It is evident from Fig. 7(a) and (c) that different p values exert a substantial impact on clustering performance, with optimal performance achieved when p is set to 0.9. Similar trends are observed in Fig. 7(b) and Fig. 7(d), with favorable performance found around p values near 0.9. This highlights the sensitivity of the Schatten p-norm to tensors, making it a powerful tool for capturing low-rank structures. Furthermore, it is noteworthy that the tensor Schatten p-norm, in comparison to the tensor nuclear norm, exhibits greater flexibility. This flexibility enables the choice of an appropriate p value for approximating low-rank tensors, thereby enhancing the overall approximation performance. (Note: When p is equal to 1, the tensor Schatten p-norm is equivalent to the tensor nuclear norm.)

I. Convergence Analyses

To verify the convergence of the proposed ETBGL, we numerically record the residual values of $\|\mathcal{B}^{k+1} - \mathcal{B}^{k}\|_{\infty}$, $\|\mathcal{E}^{k+1} - \mathcal{E}^{k}\|_{\infty}$, $\|\mathcal{S} - \mathcal{B}^{k+1} - \mathcal{E}^{k+1}\|_{\infty}$, and the objective function value in each iteration. Fig. 6 illustrates the relationship between the number of iterations and error terms for the ETBGL method across ten distinct datasets. As expected, the proposed algorithm converged to a stable state quickly. In the majority of cases, our

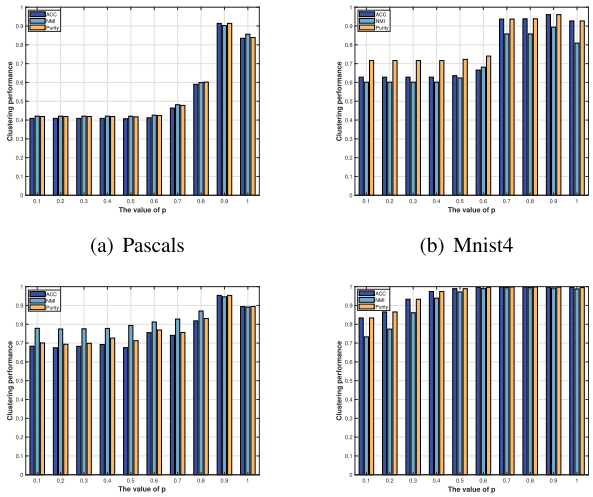
Authorized licensed use limited to: SOUTH CHINA UNIVERSITY OF TECHNOLOGY. Downloaded on September 29,2025 at 02:04:59 UTC from IEEE Xplore. Restrictions apply.



(a) BBC-Sport (b) Prokaryotic (c) Pascals (d) COIL20 (e) HW1256

(f) UCI-Digits (g) Mnist4 (h) Scene15 (i) BioData_2K3K (j) Hdigit

Fig. 6. The correspondence between the number of iterations and the error on ten different datasets.



(a) Pascals (b) Mnist4

(c) COIL20 (d) HW1256

Fig. 7. The influence of different p values on clustering performance in the ETBGL model for the Pascals, Mnist4, COIL20, and HW1256 datasets.

approach yields error values below 10^{-6} after only 15 iterations, convergence is generally reached in around 20 or 30 iterations. The model's fast convergence can also be demonstrated through the runtime shown in Table IX. This table displays the runtime of our model and other comparison methods across all experimental datasets, measured from the data loading to the generation of the final graph. It is evident that our model achieves nearly optimal runtime across all tested datasets. Notably, the comparison method TBGL lacks convergence criteria with a fixed iteration of 50 rounds. Hence, despite being a bipartite graph method, it has the longest runtime. Additionally, our model outperforms the UDBGL model, which is also a bipartite graph method. The runtime reflects the feasibility of our proposed approach on large-scale data.

VI. Conclusion

In this paper, we introduce an efficient multi-view clustering method termed Essential Tensorized Bipartite Graph Learning (ETBGL). The ETBGL model harnesses the power of the low-rank tensor Schatten p-norm to explore inter-view similarities, effectively capturing complementary information between views. Additionally, we incorporate the tensor $\ell_{2,1}$-norm as a sparse penalty term on the error component, thereby enhancing the model's robustness and preserving shared information. To efficiently optimize our approach, we present a time-economical optimization algorithm. Experimental results conducted across multiple datasets demonstrate the effectiveness of the ETBGL model.

References

[1] K. K. Sharma and A. Seal, "Multi-view spectral clustering for uncertain objects," *Inf. Sci.*, vol. 547, pp. 723–745, Feb. 2021.
[2] H. Xu, X. Zhang, W. Xia, Q. Gao, and X. Gao, "Low-rank tensor constrained co-regularized multi-view spectral clustering," *Neural Netw.*, vol. 132, pp. 245–252, Dec. 2020.
[3] B. Zhang and X. Ma, "Multi-view clustering with constructed bipartite graph in embedding space," *Knowl.-Based Syst.*, vol. 254, Oct. 2022, Art. no. 109690.
[4] Q. Gao, Z. Wan, Y. Liang, Q. Wang, Y. Liu, and L. Shao, "Multi-view projected clustering with graph learning," *Neural Netw.*, vol. 126, pp. 335–346, Jun. 2020.
[5] Y. Jia, H. Liu, J. Hou, S. Kwong, and Q. Zhang, "Multi-view spectral clustering tailored tensor low-rank representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 12, pp. 4784–4797, Dec. 2021.
[6] S. El Hajjar, F. Dornaika, and F. Abdallah, "Multi-view spectral clustering via constrained nonnegative embedding," *Inf. Fusion*, vol. 78, pp. 209–217, Feb. 2022.
[7] Y. Zhao, Y. Yun, X. Zhang, Q. Li, and Q. Gao, "Multi-view spectral clustering with adaptive graph learning and tensor schatten p-norm," *Neurocomputing*, vol. 468, pp. 257–264, Jan. 2022.
[8] K. Berahmand, M. Mohammadi, A. Faroughi, and R. Mohammadiani, "A novel method of spectral clustering in attributed networks by constructing parameter-free affinity matrix," *Cluster Comput.*, vol. 25, no. 2, pp. 869–888, Nov. 2021.

Authorized licensed use limited to: SOUTH CHINA UNIVERSITY OF TECHNOLOGY. Downloaded on September 29,2025 at 02:04:59 UTC from IEEE Xplore. Restrictions apply.

[9] D. Cai and X. Chen, "Large scale spectral clustering via landmark-based sparse representation," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1669–1680, Aug. 2015.
[10] Z. Kang, W. Zhou, Z. Zhao, J. Shao, M. Han, and Z. Xu, "Large-scale multi-view subspace clustering in linear time," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 4412–4419.
[11] X. Li, H. Zhang, R. Wang, and F. Nie, "Multiview clustering: A scalable and parameter-free bipartite graph fusion method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 330–344, Jan. 2022.
[12] L. Li and H. He, "Bipartite graph based multi-view clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 7, pp. 3111–3125, Jul. 2022.
[13] H. Chen, F. Nie, R. Wang, and X. Li, "Fast unsupervised feature selection with bipartite graph and $ell_{2,0}$-Norm constraint," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 5, pp. 4781–4793, May 2023.
[14] J. Zhu, L. Tao, H. Yang, and F. Nie, "Unsupervised optimized bipartite graph embedding," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 3, pp. 3224–3238, Mar. 2023.
[15] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, vol. 14, Art. no. 26592.
[16] A. Kumar, P. Rai, and H. Daume, "Co-regularized multi-view spectral clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, vol. 24, Art. no. 89652.
[17] W. Liu, L. Liu, L. Feng, and H. Deng, "Tensorized multi-view clustering via hyper-graph regularization," in *Proc. 2022 Int. Joint Conf. Neural Netw.*, 2022, pp. 1–8, doi: 10.1109/IJCNN55064.2022.9892515.
[18] Q. Zhou, H. Yang, and Q. Gao, "Low-rank constraint bipartite graph learning," *Neurocomputing*, vol. 511, pp. 426–436, Oct. 2022, doi: 10.1016/j.neucom.2022.09.002.
[19] R. Xia, Y. Pan, L. Du, and J. Yin, "Robust multi-view spectral clustering via low-rank and sparse decomposition," in *Proc. AAAI Conf. Artif. Intell.*, 2014, vol. 28, no. 1.
[20] M. Brbić and I. Kopriva,"Multi-view low-rank sparse subspace clustering," *Pattern Recognit.*, vol. 73, pp. 247–258, Jan. 2018.
[21] C. Zhang, Q. Hu, H. Fu, P. Zhu, and X. Cao, "Latent multi-view subspace clustering," in *Proc. 2017 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4279–4287.
[22] Y. Li, F. Nie, H. Huang, and J. Huang, "Large-scale multi-view spectral clustering via bipartite graph," in *Proc. AAAI Conf. Artif. Intell.*, 2015.
[23] S. Huang, Z. Xu, I. W. Tsang, and Z. Kang, "Auto-weighted multi-view co-clustering with bipartite graphs," *Inf. Sci.*, vol. 512, pp. 18–30, Feb. 2020.
[24] Q. Gao, P. Zhang, W. Xia, D. Xie, X. Gao, and D. Tao, "Enhanced tensor RPCA and its application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 2133–2140, Jun. 2021.
[25] H. Yang, Q. Gao, W. Xia, M. Yang, and X. Gao, "Multiview spectral clustering with bipartite graph," *IEEE Trans. Image Process.*, vol. 31, pp. 3591–3605, 2022.
[26] W. Xia, Q. Gao, Q. Wang, X. Gao, C. Ding, and D. Tao, "Tensorized bipartite graph learning for multi-view clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 5187–5202, Apr. 2023.
[27] J. Wu, Z. Lin, and H. Zha, "Essential tensor learning for multi-view spectral clustering," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5910–5922, Dec. 2019.
[28] S. Wanget al., "Fast parameter-free multi-view subspace clustering with consensus anchor guidance," *IEEE Trans. Image Process.*, vol. 31, pp. 556–568, 2022.
[29] X. Ren and Z. Lin, "Linearized alternating direction method with adaptive penalty and warm starts for fast solving transform invariant low-rank textures," *Int. J. Comput. Vis.*, vol. 104, pp. 1–14, Aug. 2013.
[30] S. Fang, D. Huang, X. Cai, C. Wang, C. He, and Y. Tang, "Efficient multi-view clustering via unified and discrete bipartite graph learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 8, pp. 11436–11447, Aug. 2024.
[31] A. Asuncion and D. Newman, *UCI Machine Learning Repository*. Irvine, CA, USA: UC Irvine, 2007.
[32] H. Wang, Y. Yang, and B. Liu, "GMC: Graph-based multi-view clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 6, pp. 1116–1129, Jun. 2020.
[33] L. Deng, "The mnist database of handwritten digit images for machine learning research [best of the web]," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 141–142, Nov. 2012.
[34] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml
[35] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 524–531.
[36] A. Farhadiet al., "Every picture tells a story: Generating sentences from images," in *Proc. Comput. Vis.–ECCV 2010, 11th Eur. Conf. Comput. Vis.*, 2010, pp. 15–29. [Online]. Available: https://doi.org/10.1007/978-3-642-15561-1_2
[37] Y. Chen, X. Xiao, C. Peng, G. Lu, and Y. Zhou, "Low-rank tensor graph learning for multi-view subspace clustering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 92–104, Jan. 2022.
[38] A. Strehl and J. Ghosh, "Cluster ensembles—A knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, 2002.
[39] Y. Zhanget al., "Collaborative embedding learning via tensor integration for multi-view clustering," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 2, pp. 1841–1852, Apr. 2024.



Wanrong Gu received the Ph.D. degree in computer system architecture from the South China University of Technology, Guangzhou, China. He was a Visiting Scholar with Washington State University, Pullman, WA, USA. He is currently a Faculty Member with the Department of Information Engineering, South China Agricultural University, Guangzhou, and master's supervisor. He is the Principal investigator of three provincial projects and one national project. He has authored or coauthored numerous papers in SCI/EI-indexed journals, including top-tier journals recommended by CCF (A/B) or ranked JCR Q1/Q2. His research focuses on software development, Big Data processing, and analysis, with research funding exceeding 1 million RMB over the past five years.



Junlong Guo is currently working toward the master's degree in computer technology with South China Agricultural University, Guangzhou, China. His research interests include machine learning, multiview clustering, and unsupervised learning.



Haiyan Wang received the Ph.D. degree in computer science and technology from the South China University of Technology, Guangzhou, China, in 2021. She is currently a Lecturer with the College of Mathematics and Informatics, South China Agricultural University, Guangzhou. Her research interests include machine learning, computer vision, and image processing.



Guangyu Zhang received the B.S. degree from Chang'an University, Xi'an, China, and the M.Sc. and Ph.D. degrees from Sun Yat-sen University, Guangzhou, China. In 2021, he joined South China Agricultural University, Guangzhou, where he is currently a Lecturer and with the Department of Computer Science. His research interests include multiview learning and applications of evolutionary computation in data mining.

Authorized licensed use limited to: SOUTH CHINA UNIVERSITY OF TECHNOLOGY. Downloaded on September 29,2025 at 02:04:59 UTC from IEEE Xplore. Restrictions apply.



Bin Zhang received the B.S. and M.S. degrees from the School of Information Science and Engineering, Shandong Normal University, Jinan, China, in 2011 and 2015, respectively, and the Ph.D. degree from the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. He is currently a Postdoctoral Researcher with the Guangdong Institute of Intelligence Science and Technology. His research interests include large-scale clustering, feature matching, cross-modal retrieval, medical image processing and bioinformatics.



Jiazhou Chen (Member, IEEE) received the Ph.D. degree in computer science and technology from the South China University of Technology (SCUT), Guangzhou, China, in 2020. During his period at SCUT, he studied with the Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, as a Visiting Scholar for one year, in 2019. He is currently an Associate Professor with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou. His research interests include bioinformatics, medical image analysis, brain network analysis, and machine learning.



Hongmin Cai (Senior Member, IEEE) received the B.S. and M.S. degrees in mathematics from the Harbin Institute of Technology, Harbin, China, in 2001 and 2003, respectively, and the Ph.D. degree in applied mathematics from The Hong Kong University, Hong Kong, in 2007. He was a Guest Professor with Kyoto University, Kyoto, Japan. He is currently a Professor with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. His research interests include bioinformatics and artificial intelligence (AI) for medical science.

Authorized licensed use limited to: SOUTH CHINA UNIVERSITY OF TECHNOLOGY. Downloaded on September 29,2025 at 02:04:59 UTC from IEEE Xplore. Restrictions apply.

SCAULIB202523388

检索证明

根据委托人提供的论文材料，委托人**华南农业大学 古万荣(学科类型:自然科学)** 1 篇论文收录情况如下表。

序号	论文名称	发表刊物及发表的年月卷期/页码等	作者排名	论文等级	作者文中单位	收录情况	影响因子	中科院大类分区
1	Efficient Multi-View Clustering via Essential Tensorized Bipartite GraphLearning	IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE 出版年：2025 卷期：9 4 页码：2952-2964 文献号： 文献类型：Article	1	A 类	South China Agricultural University	SCI	IF2-year=6.5 IF5-year=6.6 （2024）	计算机科学 2 区 Top 期刊：否 OA 期刊：否 （2025）

说明：论文等级和中科院大类分区按《华南农业大学学术论文评价方案（试行）》划分。

报告免责声明:如未盖章,报告无效

检索员：尹银怀

华南农业大学图书馆

2025-09-29

华南农业大学图书馆SCAULIB202523388

Computer Science
中国计算机学会会刊
ISSN 1002-137X
CN 50-1075/TP
CODEN JKIEBK

计算机科学

2020年 第47卷 第6A期
Vol.47 No.6A 2020



本期封面：基于多空间属性信息融合的序列推荐
王子泓, 邵蓥侠, 何吉元, 刘金宝

计 算 机 科 学

（JISUANJI KEXUE）

第 47 卷 第 6A 期 2020 年 6 月

目 次

人工智能

1 基于熵值模糊层次分析法的科技战略评价 刘子琦 郭炳晖 程 臻 等

6 基于主次关系特征的自动文摘方法 张 迎 张宜飞 王中卿 等

12 基于对抗训练的文本表示和分类算法 张晓辉 于双元 王全新 等

17 基于影响力最大化策略的抑制虚假消息传播的方法 陈晋音 张敦杰 林 翔 等

24 一种基于音高显著性增强的主旋律提取方法 金文清 韩 芳

29 基于衰减系数建立动态蛋白质网络模型进行关键蛋白质预测 戴彩艳 何 菊 胡孔法 等

34 基于可拓路径的密集人群环境通行策略 高庆吉 王闻博 侯世昊 等

40 结合百科知识与句子语义特征的关系抽取方法 吕亿林 田宏韬 高建伟 等

45 基于专利结构的中文专利摘要研究 束云峰 王中卿

49 基于船舶自动识别系统与人工神经网络的船舶载重预测 王 鹏 苏 伟 张久文 等

54 一种无约束优化的无参数填充函数算法 张玉琴 张建亮 冯向东

58 改进的支持向量回归机在电力负荷预测中的应用 唐承娥 韦 军

66 基于路口相似度的信号配时方案推荐算法 骆佳磊 孟利民

70 使用深层卷积神经网络提高 Hi-C 数据分辨率 程 哲 白 茜 张 浩 等

75 基于 sEMG 的改进 SVM+BP 肌力预测分层算法 宋 岩 胡瑢华 郭福民 等

79 基于模糊信息分解与控制规则的移动机器人沿墙导航 方梦琳 唐文兵 黄鸿云 等

84 一种基于量子耗散粒子群的评估模型构建方法 张素梅 张波涛

89 基于 ECOC 的多类代价敏感分类方法 吴崇明 王晓丹 薛爱军 等

95 基于新型文本块分割法的简历解析 祖石诚 王修来 曹 阳 等

102 基于耦合强度的多项式时间社团探测算法 杨卓璇 马源培 严 冠

108 基于区块挖掘与重组的启发式算法求解置换流水车间调度问题 陈孟辉 曹黔峰 兰彦琦

114 求解自行车共享系统静态再平衡问题的单亲遗传算法 冯炳超 吴璟莉

119 分段加权布谷鸟算法及其应用 臧 睿 刘笑笑

124 求解柔性资源受限项目调度问题的多种群遗传算法 姚 敏

130 一种新的基于函数逼近协同更新的 DQN 算法 刘青松 陈建平 傅启明 等

135 基于罗盘算子的改进蝙蝠优化算法 杨凯中 提梦桃 谢英柏

计算机图形学 & 多媒体

139 基于深度学习的行为识别算法综述 赫 磊 邵展鹏 张剑华 等

148 深度学习在光声成像中的应用现状 孙 正 王新宇
153 基于扩展 Haar 特征和 DBSCAN 的钢轨识别算法 罗晋楠 张济民
157 融合多颜色空间信息的彩色图像色差校正算法 唐佳林 张 冲 郭岩锋 等
161 全局双边网络的语义分割算法 任天赐 黄向生 丁伟利 等
166 多尺度膨胀卷积在图像分类中的应用 吴昊昊 王方石
172 基于卷积神经网络的半监督手术视频流程识别 祁宝莲 钟坤华 陈芋文
176 基于改进 YOLO-V2 算法的遥感图像目标检测技术研究 张 曼 李 杰 丁荣莉 等
181 基于多尺度自适应权重的稀疏表示目标跟踪算法 程中建 周双娥 李 康
187 结合 EHHT 和 CI 的精神分裂症语音自动检测算法 田维维 周 悦 尹 旺 等
196 动态多特征融合的道路遗洒物威胁度分析方法 吴宏涛 刘力源 孟 颖 等
206 基于主动轮廓演变模型的遥感影像单棵树木检测 叶 阳 周棋正 沈 瑛 等
213 基于卷积神经网络的污损遮挡号牌分类 李 林 赵凯月 赵晓永 等
220 基于改进树种算法的彩色图像多阈值分割 彭 浩 和丽芳
226 一种基图像提取和内容无关图像重构方法研究 蓝章礼 申德兴 曹 娟 等
230 基于残差生成对抗网络的人脸图像复原 李泽文 李子铭 费天禄 等
237 基于视觉显著性的海面船只候选区域检测方法 刘俊琦 李 智 张学阳
242 基于 BP 神经网络的摔倒检测算法 周立鹏 孟利民 周 磊 等
247 基于直方图差异的工业产品表面缺陷检测方法 杨志伟 戴 铭 周智恒
250 三角坐标系下人脸表情表示方法 肖 潇 孔凡芝
254 用于肺结节影像分类识别的 DBN 与 CNN 的比较研究 张华丽 康晓东 冉 华 等
260 基于多模型优化的超声图像肿瘤自动识别 古万荣 樊纬江 谢贤芬 等

计算机网络

268 边缘计算技术及应用综述 赵 明
273 基于 CQPSO 移动群智感知紧急任务分配方法研究 李建军 汪校铃 杨 玉 等
278 一种基于能量补给的无线传感器网络分簇单跳路由协议 冯 钧 孔建寿 王 刚
283 低信噪比下基于深度学习的调制模式识别方法 陈晋音 成凯回 郑海斌
289 基于 AMI 的智慧用能多模汇聚网关装置研究 肖 勇 金 鑫 王立博 等
294 一种融合 EMD 分解和 LSTM 网络的频谱占用度预测模型 赵晓东 苏公瑾 李克利 等
299 互连网络的模 p 剩余类加群的笛卡尔积模型 师 腾 师海忠
305 一种参数可变的最小存储再生码 王雪冰
310 基于信噪比分级的信号调制类型识别 陈晋音 蒋 焘 郑海斌
318 面向云服务的分布式消息系统负载均衡策略 高子妍 王 勇

信息安全

325 图像隐写分析算法研究概述 彭 伟 胡 宁 胡璟璟
332 区块链共识算法对比研究 陆歌皓 谢莉红 李析禹
340 网络安全态势感知研究现状与发展趋势的图谱分析 白 雪 努尔布力 王亚东

344 三次 MI 多变量公钥密码体制的安全性分析 张 栖 聂旭云
349 基于布尔导数的布尔置换构造 吴万青 周国龙 马晓雪
352 基于改进 BP 神经网络的 SQL 注入识别 诸珺文
360 工业互联网网络传输安全问题研究 伍育红 胡向东
364 基于用户特征提取的新浪微博异常用户检测方法 袁得嵛 章逸钒 高 见 等
369 基于区块链的制造联盟系统 洪小玲 万 虎 肖 晓 等
375 针对 AES 查表法最后一轮加密的 L3 缓存攻击 陆 垚 陈开颜 王寅龙 等
381 基于 C4.5 决策树的 HTTPS 加密流量分类方法 邹 洁 朱国胜 祁小云 等
386 一种基于环签名和短签名的可净化签名方案 张君何 周清雷 韩英杰
391 基于 Gossip 协议的信任收集共识算法研究 张奇文 王志强 张逸谦
395 面向缺损数据的(α,k)-匿名模型 张王策 范 菁 王渤茹 等
400 基于 TFR 模型的公安云平台数据分级分类安全访问控制模型研究 顾荣杰 吴治平 石 焕
404 基于软件定义网络资源优化的虚拟网络功能部署策略 黄梅根 汪 涛 刘 亮 等
409 PTN 网络安全结构的优化与设计 闫 震 田 毅 段志国 等
413 一种基于无证书的 SIP 认证密钥协商协议 莫天庆 何咏梅
420 大属性可公开追踪的密文策略属性基加密方案 马潇潇 黄 艳

数据库 & 大数据 & 数据科学

424 电商平台用户再购物行为的预测研究 吕泽宇 李纪旋 陈如剑 等
429 改进的 GHSOM 算法在民航航空法规知识地图构建中的应用 张浩洋 周 良
436 基于面向对象(属性)概念格的形式背景属性约简方法 岳晓威 彭 莎 秦克云
440 一种基于时序性告警的新型聚类算法 邓甜甜 熊荫乔 何贤浩
444 基于 CEEMD-Pearson 和深度 LSTM 混合模型的 $PM_{2.5}$ 浓度预测方法 丁子昂 乐曹伟 吴玲玲 等
450 基于判断聚合的分布式数据挖掘分类算法研究 李 莉
457 基于 Xie-Beni 指数的选择性聚类集成 邵 超 马进家
461 基于谱聚类的多目标进化社区发现算法研究 董明刚 弓佳明 敬 超
467 基于 LSTM-GA 的股票价格涨跌预测模型 包振山 郭俊南 谢 源 等
474 基于机器学习的 HBase 配置参数优化研究 徐江峰 谭玉龙
480 改进的局部和相似性保持特征选择算法 李金霞 赵志刚 李 强 等
485 基于 GM(1,1)-SVM 组合模型的中长期人口预测研究 徐翔燕 侯瑞环
488 一种新的不均衡关联分类算法 崔 巍 贾晓琳 樊帅帅 等
494 基于牛顿法的自适应高阶评分距离推荐模型研究 邹海涛 郑 尚 王 琦 等
500 基于改进残差网络的水下图像重建方法 宋娅菲 谌雨章 沈君凤 等
505 一种基于张量的车辆交通数据缺失估计新方法 张德干 范洪瑞 龚倡乐 等
512 基于 X12-LSTM 模型的保费收入预测研究 刁 莉 王 宁
517 基于 GA-SVM 的农产品质量分类 马 创 吕孝飞 梁炎明
521 基于用户兴趣的农产品推荐技术研究 李建军 付 佳 杨 玉 等
526 覆盖近似空间中的核及其性质 周俊丽 管延勇 徐法升 等

530 基于文本挖掘和决策树分析的中国手游产业发展研究 朱涤尘 夏 换 杨秀璋 等
535 基于灰色——马尔可夫模型的农产品产量预测方法 马 创 袁 野 尤海生
540 基于卷积神经网络与约束概率矩阵分解的推荐算法 马海江

交叉 & 应用

546 基于 Hyperledger Fabric 的汽车供应链系统的方案设计与性能分析 林旭丹 鲍士兼 赵立昕 等
552 基于分支定界法的机场终端区单一进离场程序设计优化 周 隽 王天淇
556 私有二进制协议中变长域的格式挖掘方法 徐旭东 张志祥 张 献
561 基于 LLVM 编译器的节点融合优化方法 胡 浩 沈 莉 周清雷 等
567 系统仿真可信度评估方法研究 郭聪蕊 王 珺 封一鸣
572 基于图论的组织互操作性建模与评估研究 高 琳 段国林 姚 涛
577 切换拓扑条件下的多无人机系统编队包含控制 赵学远 周绍磊 王帅磊 等
583 一种替代性纠纷在线仲裁系统 周 蔚 罗旭东
591 区块链技术在信托行业的应用研究 可雨憬 敬茂华 郑涵尹
596 基于 6LoWPAN 的低功耗长距离海洋环境监测系统 王 栋 王 虎 姜迁里
599 变电站巡检机器人重定位研究 李忠发 杨 光 马 磊 等
603 基于区块链技术的通证模型的设计与分析 巫光福 陈 颖 曾宪文 等
609 基于 FPGA 的智能视频加速检索系统 赵 博 杨 明 汤志伟 等
612 基于 BigQuant 大数据平台的股票投资策略开发 李 泳
616 基于直觉模糊层次分析的特种设备事故应急预案评价 郑耿峰
622 基于嵌入式多核 DSP 的加速软件系统 蔡玉鑫 汤志伟 赵 博 等
626 基于粗糙规则的脉冲神经膜系统计算能力的研究 罗云芳 唐承娥 韦 军
631 基于 5G 的视觉辅助 BDS 移动机器人融合定位算法 马 虹
634 基于区块链与人工智能的网络多服务器 SIP 信息加密系统设计 任 仪
639 基于区块链构建新型企业客户服务技术平台 张启明 陆建华 李守智 等
643 基于微服务架构的 ETC 系统设计 于 曼 黄 凯 张 翔
648 嵌入式设备的邮递式升级方案 陈 云

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191200011

基于多模型优化的超声图像肿瘤自动识别

古万荣[1]　樊纬江[1]　谢贤芬[2]　张子烨[3]　毛宜军[1]　梁早清[1]　林镇溪[1]
1 华南农业大学数学与信息学院　广州 510642
2 暨南大学经济学院　广州 510632
3 华南理工大学数学学院　广州 510641

摘　要　随着计算机视觉识别技术的发展,越来越多的研究人员将该技术应用到肿瘤图像的识别上。但由于成本,许多医院仍然采用成本较低的B超等设备,产生了模糊、伪影和多个相似肿瘤噪声区域。目前的方法在清晰图像识别中具有很高的精度,但在超声图像方面却存在低准确度且不稳定的结果,其原因是许多现有算法对模糊、噪声图像误判较高。文中基于R-CNN和PRN的方法快速准确地获取高噪声的超声图像的关键特征,并通过数据增强和形态学滤波的方法确保了识别的稳定性。同时,所提方法还融合了血流信号分类模型增强识别精度。基于实际甲状腺肿瘤图像的数据集可知,提出的方法对比新近算法具有较高的准确性和稳定性。

关键词:深度学习;融合模型;肿瘤识别;神经网络
中图法分类号　TN957.52;TP751

Automatic Tumor Recognition in Ultrasound Images Based on Multi-model Optimization

GU Wan-rong[1],FAN Wei-jiang[1],XIE Xian-fen[2],ZHANG Zi-ye[3],MAO Yi-jun[1],LIANG Zao-qing[1] and LIN Zhen-xi[1]
1 School of Mathematics and Information,South China Agricultural University,Guangzhou 510642,China
2 School of Economy,Jinan University,Guangzhou 510632,China
3 School of Mathematical,South China University of Technology,Guangzhou 510641,China

Abstract　With the development of computer vision recognition technology,more and more researchers apply this technology to the recognition of tumor images. But because of the cost,many hospitals still use low-cost ultrasound and other equipment,resulting in ambiguity,artifacts and many similar tumor noise areas. The present method has high precision in clear image recognition,but it shows low accuracy and unstable result in ultrasonic image. The reason is that many existing algorithms misjudge the modulus and noise image. In this paper,the key features of high-noise ultrasound images are obtained quickly and accurately by R-CNN and PRN methods,and the stability of recognition is ensured by data enhancement and morphological filtering. At the same time,the classification model of blood flow signal is fused to improve the recognition accuracy. Based on the data set of a real Thyroid neoplasm image,the proposed method is more accurate and stable than the new algorithm.

Keywords　Deep learning,Fusion model,Tumor recognition,Neural network

多普勒超声是一种常用的医学诊断技术。与其他成像技术相比,其具有实时成像、成本低、无需活检、对患者无损伤等优点。尽管有这些优点,其在超声图像的肿瘤识别方面仍然存在一些问题,如严重斑点、高噪声、低分辨率、对比度不明显以及肿瘤形状无规则。上述缺点导致了肿瘤自动识别过程的准确性和不稳定性的降低。为了克服多噪声图像中的肿瘤识别问题,近年来有许多研究成果。但大多数研究集中在降噪和图像锐化方面,而不是将两者结合起来进行分析,以适应特定的症状和生理特征,以达到适当的水平。此外,目前大多数研究仍然采用传统的特征选择方法,如互信息量和主成分分析(PCA),有效信息大量丢失。目前人们已经提出了深度学习方法来进行特征挖掘,并且已经取得了一些成果,但是一些特征提取算法在医学图像处理中并不能很好地应用。特别是它们不适合肿瘤超声图像,准确性和稳定性仍然不足。在现实的医疗中,高误诊率会延误医疗的进行。

在以上背景下,肿瘤图像识别研究至少存在两个主要问题:1)输入数据为低质量的超声图像,换句话说,图像是模糊的,包含很多噪声或模糊的形状边缘;2)主流的特征选择技术,如深度学习和机器学习,对特定的真实世界的甲状腺超声图像适应性较差。为了解决这两个问题,本文提出采用数据

基金项目:广东省自然科学基金(2018A030313437);广东省哲学社会科学项目(GD18CXW01);广东省科技计划项目(2018A070712021);教育部人文社科项目(18YJCZH037);全国统计科学研究重点项目(2019LZ37)
This work was supported by the Guangdong Natural Science Foundation Project (2018A030313437),13th Five-year Plan Project of Philosophy and Social Science in Guangdong Province (GD18CXW01),Guangdong Science and Technology Program Project (2018A070712021),Ministry of Education Humanities and Social Sciences Research Youth Fund Project (18YJCZH037) and 2019 National Statistical Science Research Key Project(2019LZ37).
通信作者:毛宜军(yijunmao@163.com)

增强、形态滤波等方法保证数据质量，有效地利用数据质量进行特征提取，然后利用 R-CNN 和 RPN 多层深度网络充分挖掘图像海量的潜在特征，这样就可以挖掘出肿瘤区域和噪声区域之间的差异。同时，该算法可以更准确地锁定肿瘤区域，降低误报率。由于该方法的高精度和高效率，肿瘤区域可以被快速锁定，未来可以将其应用于动态视频甲状腺癌的识别中。

1 相关技术背景

医学图像识别一般包括 3 个主要步骤：图像预处理、特征抽取和对象识别。

一般来说，输入的图像在进行图像识别之前需要进行去噪、对比度增强、锐化、区域分割等处理。本文中，超声图像承载着大量的噪声信息，图像的模型和预处理是必不可少的。Yu-Len 等[2]提出了一种乳腺肿瘤的分水岭分割方法，Gaetano 等[3]对遥感图像的分割进行了研究。本文主要研究医学图像的去噪问题。一些研究建议对图像分割进行特征缩减。越来越多的研究集中在模糊医学图像的分割上[6]。然而，现有的预处理方法在清晰彩色图像处理方面有很好的效果，并且在许多应用领域都有很好的性能[7-8]。甲状腺癌图像只有黑白模具和边缘，有更多的脉冲噪声干扰，因此现有的方法大多不适合本文的应用。

本文基于形态学和纹理特征，从图像中识别良恶性结节外观，从而实现识别算法[9]。其形态特征包括结节的方向、形状、回声、边缘等，其在许多应用中得到应用[10-11]。纹理特征包括超声表现和回波模型。传统的特征提取方法主要基于形状[12]、颜色[13]和纹理[12,14]，其常用的方法有 SIFT 和 HOG[15]。

分类器或聚类器用于识别病变区域的分割图像。在进行分类之前，需要筛选出分类性能较强的特征，以提高分类器的泛化能力，并降低计算维数，避免维数灾难。由于良恶性肿瘤的特征具有自相关性和交互性，非线性映射方法可以作为分类器或聚类器。常用的方法有 BP 算法、决策树算法、FLD 算法、SVM 算法、贝叶斯分类器算法、SOM 算法等。然而，传统的机器学习特征提取方法过于依赖经验。在许多新的应用程序场景中，没有太多可以学习的经验。因此，这类场景的特征提取存在较大的误差。

为了解决传统机器学习方法在特征挖掘中的不足，本文提出了一种新的深度学习方法，将深度学习算法应用到医学图像识别和图像分割分析中。在医学图像分割领域，不同的分割方法通常与不同的机器学习或深度学习方法相结合，如 Mocanu 等提出的基于特征学习和集成学习分层视网膜血管分割的方法[16]，该方法也可用于跟踪心室内膜超声图像分割。此外，3D CNN 和全卷积神经网络常用于语义分割和图像分割。例如 Korez 等利用 3D CNN 对 MR 图像进行脊柱分割，Moeskops 等在 MRI 图像中对大脑和胸部肌肉进行分割。在分类和识别领域，有监测视网膜技术的糖尿病患者的转移学习等[19]、皮肤癌的分类转移学习[20]、DCE-MRI 的自动分类稀疏自动编码器[21]和检测和评价膝骨关节炎分级使用的卷积神经网络[22]等方法。随着计算机视觉技术的不断发展，深度学习模型在辅助诊断方面也取得了很大的进展，如利用 Bengtsson 等[23]的 Deep Belief 网络和 SVM 对 pap-spar 进行分析，对宫颈癌进行识别；Kooi 等[24]基于 CNN 对乳腺良恶性肿瘤进行识别。这些方法远远优于传统的计算机辅助诊断方法。此外，深度学习算法常用于判断前列腺疾病、乳腺癌、肺癌以及研究细胞微核检测等。

然而，大多数甲状腺肿瘤的超声图像分割方法都不能实现端到端的分割，这就需要人工标注干预。首先，感兴趣区域(ROI)是由具有丰富临床经验的医生在原始超声图像标记。然后利用各种边缘分割算法提取感兴趣区域的轮廓，这就需要大量的人力和精力，需要医生具有较高的临床经验和个人能力。因此，如何能高效且高精度地实现肿瘤感兴趣区域的自动识别具有重要意义。

2 多模型融合的方法

目前深度学习图像识别中的目标检测方法主要有两种：1)基于感兴趣区域提取的两阶段模型，如 R-CNN，SPP-Net，Fast R-CNN，Faster R-CNN，R-FCN 等；2)不提取 ROI 的一阶段模型，如 YOLO，SSD 等。作为计算机视觉领域的重要组成部分，这些图像识别方法主要是利用细胞神经网络进行特征学习，结合一些传统的视觉方法实现目标检测。

超声图像的成像原理是利用超声波穿透人体器官，产生黑白图像。一般来说，噪声区域不仅模糊而且易受干扰，容易形成噪声区域。图 1 给出甲状腺肿瘤超声图像示例。



图 1 甲状腺肿瘤的超声图像

Fig. 1 Ultrasound image of thyroid neoplasm

为了解决甲状腺癌的图像识别问题，本文提出的基本技术框架如图 2 所示。

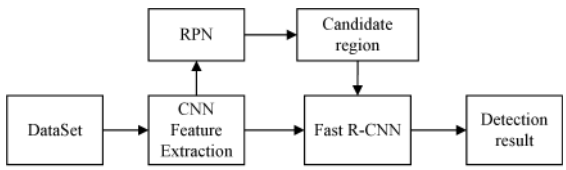


图 2 基于 Fast R-CNN 的甲状腺肿瘤图像识别模型结构

Fig. 2 Structure of thyroid neoplasm image recognition model based on Fast R-CNN

基于快速 R-CNN 的思想，本文提出分别构造 RPN 和快速 R-CNN 来提取和检测图像中的区域。采用不同的卷积网络如 ZF，VGG，ResNet50 等进行特征提取。最有效的网络将用作 RPN 和 Fast R-CNN 的共享网络。将 RPN 生成的候选区域输入到 Fast R-CNN 中，用于软最大分类和边界回归。

图 3 展示了卷积神经网络的基本结构，主要包括输入层(原始超声图像)、卷积层、池层、全连接层和分类层(软最大层)。

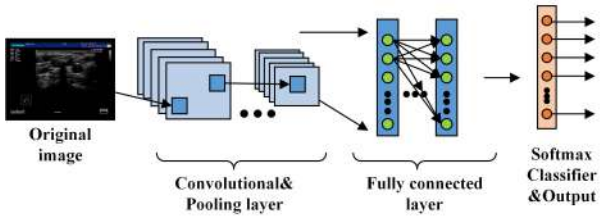


图 3 卷积神经网络基本结构

Fig. 3 Structure of convolutional neural network

1)卷积层:卷积层是卷积神经网络的重要组成部分,负责卷积神经网络的许多计算任务。每个卷积层包含多个卷积核(滤波器),通过平移和滑动扫描整个图像。此外,这些卷积核可以通过向前传播和向后传播进行自学习。通过卷积核进行图像扫描的结果称为特征图或特征映射(特征映射),表示每个空间位置对卷积核输入的响应。公式如下:

$$x_j^l = f\left(\sum_{i\in M_j} x_i^{l-1} * w_{ij}^l + b_j^l\right) \tag{1}$$

以二维图像 i 作为输入,使用二维卷积核 k 进行卷积运算后得到的特征如下:

$$\begin{aligned}\boldsymbol{S}(i,j) &= (\boldsymbol{I} * \boldsymbol{K})(i,j)\\ &= \sum_m \sum^n \boldsymbol{I}(m,n)\boldsymbol{K}(i-m, j-n)\end{aligned} \tag{2}$$

一般情况下,卷积层的结构可由卷积核的大小、卷积核的个数、步长和填充 4 个超参数确定。不同的卷积核可以提取不同的特征。步长决定了卷积核每个运动的长度,填充可以提取图像的边缘信息,这基于两个原则:局部感知和参数共享。

2)ROI 池化层:ROI 池化层使用最大池化操作将 ROI 特征转化为维度较少并且固定的特征图,大小为 $H\times W$(如 7×7),其中 H 和 W 都是超参数,需要事先设置。每个 ROI 由 4 个参数决定(r,c,h,w),其中(r,c)为左上角坐标,(h,w)为高和宽。ROI 可以利用 $h/H\times w/W$ 大小的尺寸作为滑动窗口,将大小为 $h\times w$ 的 ROI 转化为 $H\times W$ 的特征图,从而使得 POI 池化层能够处理任意尺度的特征图。

3)损失函数:Fast R-CNN 同样也有两个损失,分别为分类损失和回归损失,这种多任务损失通常将所有的损失叠加起来,从一个损失权重衡量参数来表示。其损失函数定义为:

$$L(p,u,t^u,v) = L_{cls}(p,u) + \lambda[u\geqslant 1]L_{loc}(t^u,v) \tag{3}$$

其中,$p=(p_0,p_1,\cdots,p_K)$表示每个类别的预测概率;K 为类别个数,一共包括 $K+1$ 个类别,第 0 个类别为背景;u 为真实边框的类别,$t^u=(t_x^u,t_y^u,t_w^u,t_h^u)$为第 u 个类别的预测边框的坐标,$v=(v_x,v_y,v_w,v_h)$为第 u 类的真实边框的坐标;$[u\geqslant 1]$是指示符函数,当 $u\geqslant 1$ 时其值为 1,否则为 0;λ 为权重衡量参数;L_{cls} 为类别的损失函数,L_{loc} 为边框回归损失函数,公式如下:

$$L_{cls}(p,u) = -\log p_u \tag{4}$$

$$L_{loc}(t^u,v) = \sum_{i\in\{x,y,w,h\}} smooth_{L1}(t_i^u - v_i) \tag{5}$$

本文采用可以自学习的 RPN 网络生成 ROI,如图 4 所示,RPN 最终有两个输出,因此该网络可以看成一个二类分类器。由于每一个 Anchor 输出一个二类分类值,因此使用了 k 个 Anchor,最终输出的是长度为 $2k$ 的向量。而另一个为回归输出,每个 Anchor 对应一个长度为 4 的向量(x,y,w,h),其中(x,y)为 Anchor 的中心坐标,(w,h)为 Anchor 的长和宽,则一共包含 $4k$ 个输出。

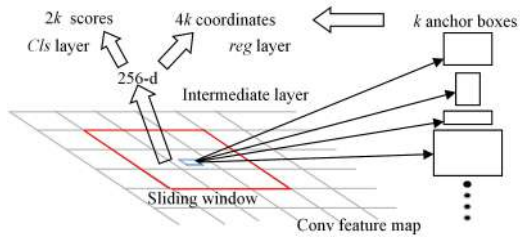


图 4 RPN 网络结构图

Fig. 4 RPN network structure

根据上面的定义,RPN 的多任务损失函数(Loss Function)见式(6):

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_t L_{cls}(p_i,p_i^*) + \lambda\frac{1}{N_{reg}}\sum_t p_i^* L_{reg}(t_i, t_i^*) \tag{6}$$

其中,i 为每个批量 Anchor 的下标,p_i 为第 i 个 Anchor 是目标类的预测概率,如果 p_i^* 是目标类别,则 p_i^* 为 1,否则为 0;t_i 是一个长度为 4 的向量,表示预测的边框的参数化坐标,t_i^* 是一个长度为 4 的向量,表示真实的边框的坐标;N_{cls} 为小批量采样对的大小,N_{reg} 为 Anchor 的数量;λ 为平衡损失函数的权重,通常设为 10;L_{cls} 为类别的损失函数,L_{reg} 为回归损失函数。

$$L_{cls}(p_i,p_i^*) = -\log[p_i^* p_i + (1-p_i)(1-p_i^*)] \tag{7}$$

$$L_{reg}(t_i,t_i^*) = smooth_{L1}(t_i - t_i^*) \tag{8}$$

$$smooth_{L1}(x) = \begin{cases} 0.5x^2, & \text{if } |x|<1 \\ |x|-0.5, & \text{if } |x|\geqslant 1 \end{cases} \tag{9}$$

损失函数中当 $p_i^*=1$ 时才有损失,否则不计算损失。t_i 和 t_i^* 的计算见式(10)-式(17):

$$t_x = (x-x_a)/w_a \tag{10}$$

$$t_y = (y-y_a)/h_a \tag{11}$$

$$t_w = \log(w/w_a) \tag{12}$$

$$t_h = \log(h/h_a) \tag{13}$$

$$t_x^* = (x^*-x_a)/w_a \tag{14}$$

$$t_y^* = (y^*-y_a)/h_a \tag{15}$$

$$t_w^* = \log(w^*/w_a) \tag{16}$$

$$t_h^* = \log(h^*/h_a) \tag{17}$$

其中,x,y,w,h 分别是预测边框的中心坐标,以及长和宽;x^*,y^*,w^*,h^* 分别是真实边框的中心坐标,以及长和宽;x_a,y_a,w_a,h_a 分别是 Anchor 边框的中心坐标,以及长和宽。

4)训练方式:RPN 的训练可以联合后向传播和 SGD 实现端到端训练,每次从一张图像中抽取小批量的样本,每批样本包括正负样本,由于负样本的数量远远大于正样本的数量,所以在采样时可以偏向负样本,例如按照 1:1 的正负比例采样 256 个样本,如果正样本不满足 128 个,则可以使用负样本进行填充。

RPN 的一个卷积层可以使用标准差为 0.01 均值为 0 的高斯分布初始化,而其他的层可以用已经在大规模图像数据库(Pascal VOC, ImageNet 等)中训练好的预训练模型(如 ZF, VGG, ResNet)初始化,然后进行微调。

用预训练模型分别初始化 RPN 和 Fast R-CNN 网络,接着对 RPN 进行训练,训练完生成 ROI,然后用这些 ROI 训练 Fast R-CNN 网络,再用 Fast R-CNN 网络初始化 RPN,与一开始初始化 RPN 不同的是,这里固定共享卷积层的参数,只是微调 RPN 的参数,接着对 RPN 训练来初始化 Fast R-CNN 网络,使得 RPN 和 Fast R-CNN 网络能够交替训练。这种训

练方式效果比较好，但是耗时较长。

这种训练方式尝试把 RPN 和 Fast R-CNN 网络融合到一个网络里，在前向传播时 Fast R-CNN 网络使用 RPN 产生的 ROI 进行训练，在反向传播时共享层的梯度来源于 RPN 和 Fast R-CNN 网络。这种情况忽略了边框坐标的梯度，是一种近似的训练，但能够减少 25%～50%的训练时间。

本文中，RPN 网络损失函数的参数设置为：$N_{cls}=256$，$N_{reg}=2400$，$\lambda=10$。ROI 池化层参数设置为：$H\times W=7\times 7$。Fast R-CNN 网络损失函数的参数设置为：$\lambda=1$。

3 实验与分析

本文所有的甲状腺结节超声图像均来自中山大学附属肿瘤医院头颈门诊以及头颈二区科。其中包括 16 个病例，368 幅超声图像，包括良恶性图像。但是由于未标注出肿瘤的良恶性，所以本文只考虑肿瘤位置的识别，不包括良恶性的识别。实验将 368 幅超声图像按照 5∶2∶3 的比例划分为训练集、验证集以及测试集。

在实验中使用近似联合训练方法对模型进行训练，采用基于动量的随机梯度下降优化算法，其梯度更新方式为：

$$v_t=\gamma v_{t-1}+\eta\times g(\theta)$$

$$\theta=\theta-v_t$$

其中，θ 为需要更新梯度的参数；v_t 为 t 时刻的动量；γ 是动量参数且 $\gamma\in[0,1]$；η 为学习率；$g(\theta)$ 为 θ 的梯度。在实验中，将 γ 设置为 0.9，初始学习率 η 设置为 0.001，每迭代 50 000 次衰减 0.0005。

3.1 不同模型的准确度对比

根据上述的实验设置，我们分别采用 ZF，VGG16，ResNet50 3 种模型作为基线方法，并以不同的迭代次数对模型进行训练。由于实验数据集数据量不大，所以均采用网络在 ImageNet 上的预训练模型，然后进行微调。实验结果对应的 *mAP*（Mean Average Precision）如表 1 和图 5 所示。

表 1 不同迭代次数以及不同网络模型的 *mAP*

Table 1 *mAP* for different iterations and different network models

（单位：%）

网络模型	迭代次数		
	1 000	5 000	10 000
ZF	0.691	0.762	0.697
VGG16	0.732	0.786	0.786
ResNet50	0.720	0.799	0.772

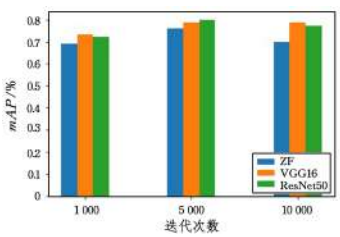


图 5 不同迭代次数以及不同网络模型的 *mAP*

Fig. 5 *mAP* histogram with different iterations and different network models

从表 1 和图 5 中可以获取的主要信息有：

（1）VGG16 和 ResNet50 的效果要比 ZF 好，主要原因是 ZF 为小型网络，只有 7 层，而 VGG16 和 ResNet50 均为大型网络，层数达到了 16 层（VGG16）或者 50 层（ResNet50）。通过增加网络层数可以使得网络提取到更深层语义更丰富的信息，从而提高模型的准确率。

（2）迭代次数为 5 000 时效果最好，而迭代次数为 10 000 次出现了下降的情况，主要原因是迭代 10 000 次时模型出现了过拟合，而迭代 1 000 次时欠拟合。由于实验数据量太少，所以整个模型很容易收敛，迭代次数不宜过多。

图 6 显示了 3 种网络模型的 *Loss* 随着迭代次数的变化折线图。可以发现，3 种网络均在 3 000 次迭代时基本收敛，随后基本趋于平稳状态，其中 VGG16 会出现较大的起伏，而 ResNet50 效果最好，不仅 Loss 最小，而且也最平稳。所以，后期的系统我们使用 ResNet50 作为 Faster R-CNN 中的特征提取网络。图 6 也可以指导后期网络模型的训练次数，以便减少时间开销和防止过拟合。

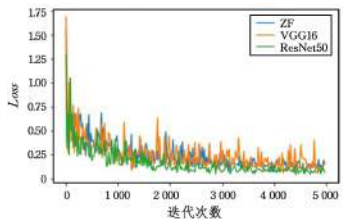


图 6 不同网络模型的 *Loss* 折线图

Fig. 6 Polyline diagrams of *Loss* for different network models

3.2 实验分析和调整

根据表 1 可知，3 种模型最好的结果只有 79.9%，还有改进的空间。由于数据集较小，很多因素都会影响实验结果，如数据集划分比例、迭代次数、学习率、数据增强等。下面主要从数据集划分比例、数据增强、学习率因素改进本文模型。

（1）数据集划分比例

由于数据集太小，很容易收敛，不同的划分会使得训练集和测试集比例不同，对模型的结果也会产生影响。如果训练集比测试集大得多，则可能出现过拟合；如果训练集比测试集小或者相差不大，则可能出现欠拟合。因此，我们将迭代次数设置为 3 000，为了减少训练时间使用 ZF 网络进行测试，采用 4 种不同的划分比例，分别为 2.5∶2.5∶5，5∶2∶3，4∶4∶2，4∶3∶3。实验结果如表 2 和图 7 所示。

表 2 不同的数据划分比例下的 *mAP*

Table 2 *mAP* of different data partition ratios

（单位：%）

网络模型	划分比例（训练集∶验证集∶测试集）			
	2.5∶2.5∶5	5∶2∶3	4∶4∶2	4∶3∶3
ZF	0.775	0.762	0.848	0.778

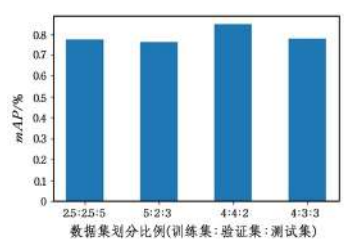


图 7 不同数据集划分比例下的 *mAP* 柱状图

Fig. 7 *mAP* histogram with different data set partition ratios

从表 2 和图 8 可以看出，不同的数据集划分比例的确会影响模型的最终效果。可以看出，使用 4∶4∶2 的划分比例时准确率从 76.2%（5∶2∶3）提高到了 84.8%。因为数据量小，将更多的数据用于训练更有助于模型学习更多的特征，而且

训练集和验证集的比例比较均匀，能够防止模型过拟合，从而提高模型的泛化能力。这 4 种划分比例的 *Loss* 变化折线图如图 8 所示，可以看出不同的划分比例对收敛速度影响不大，基本都是在迭代 3000 次时基本收敛，达到稳定状态。

从表 2 和图 8 可知，数据集太小，如果用于训练的数据太少使得模型学习到的特征太少，则不容易表达出具体的语义信息；而用于训练的数据太多测试数据太少则使得模型很容易过拟合，对测试数据的表达能力不强，泛化能力弱。为了提高模型的泛化能力，我们使用数据增强技术，例如水平翻转、垂直翻转、随机截取、固定截取、尺度变换、旋转、色彩抖动、边缘增强等。有效的数据增强不仅能扩充训练样本的数量，还能增强样本的多样性，一方面可以避免过拟合，另一方面会带来模型性能的提升，增加图像的平移不变性和旋转不变性，增加卷积神经网络对物体尺度和方向上的鲁棒性。

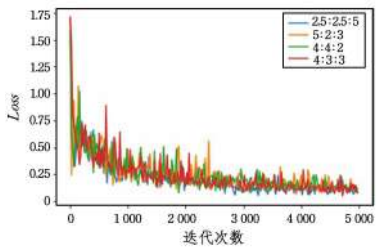


图 8 数据集不同划分比例的 *Loss* 变化折线图

Fig. 8 Polyline diagram of *Loss* variations for different partition proportions of the dataset

LeNet 在测试时将测试图像归一化，按照固定大小截取 4 个角以及中心一共 5 个区域，并将图片水平翻转，增加图像的旋转不变性和平移不变性，提高模型的泛化性能。经过这些操作后，图像数据集可以扩充为原来的 10 倍，然后用投票机制选择最合适的类别。本文借鉴这种思想，采用随机截取再加上水平翻转的方式增加数据集，其中随机截取的大小为 224×224。除此之外，还使用图像处理中的边缘增强增加超声图像的对比度。本文使用上述方法对数据集进行数据增强，在 4 种不同的数据集划分比例中采用 ZF 网络做了实验，实验结果如表 3 和图 9 所示。

表 3 数据增强在不同数据集划分下的 *mAP* 比较

Table 3 Data enhancement *mAP* comparison under different data set partition

（单位：%）

是否进行数据增强	划分比例(训练集:验证集:测试集)			
	2.5:2.5:5	5:2:3	4:4:2	4:3:3
否	0.775	0.782	0.848	0.778
是	0.862	0.902	0.802	0.889

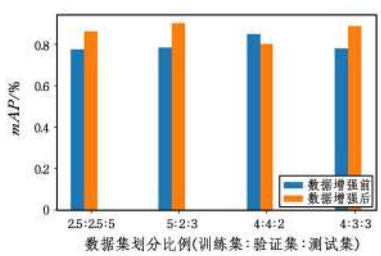


图 9 不同数据集划分比例下数据增强前后的 *mAP*

Fig. 9 *mAP* histogram before and after scaling of different datasets

从实验结果可知，使用数据增强后准确率普遍上升，4 种划分比例的准确率都在 80%以上，其中划分比例为 5:2:3 的准确率从原先的 78%提高到了 90%，而未使用数据增强的准确率基本都在 70%～80%。使用数据增强后，不仅扩充了样本的数量，同时增加了图像的平移不变性和旋转不变性，提高了模型的鲁棒性和准确度。除此之外，本文对 4 种划分比例下使用数据增强后的 *Loss* 收敛变化做了分析，从图 10 可以看出，数据增强后 4 种不同划分比例下的 *Loss* 收敛速度基本保持一致，同样在迭代 3000 次左右达到稳定状态。图 11 为不同划分比例下数据增强前后的 *Loss* 变化，4 种划分比例下数据增强前后的 *Loss* 收敛速度也基本不变。

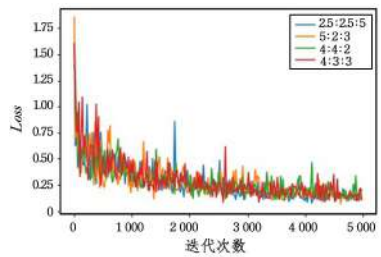


图 10 数据集增强前后数据集不同划分比例下的 *Loss* 变化折线图

Fig. 10 Polyline diagram of *Loss* changes before and after data set enhancement with different partition ratios

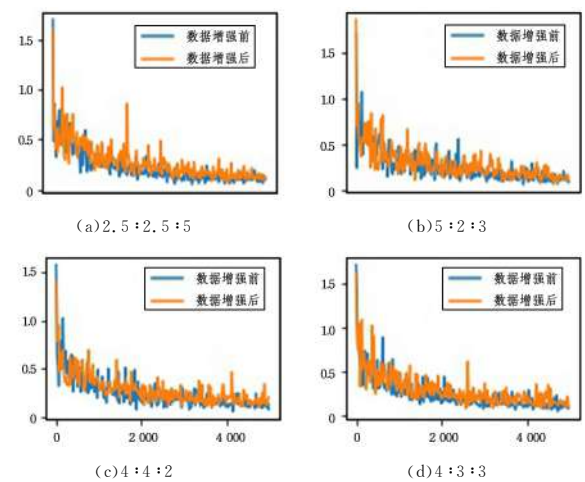


图 11 数据增强前后不同划分集的 *Loss* 变化趋势

Fig. 11 Trends in *Loss* changes for different partition sets before and after data enhancement

前文采用随机截取的方式做数据增强，但是其缺点是截取出的信息很有可能不完整，如图 12 所示，所截取的肿瘤只有一半的区域，属于无效信息，这会对模型的识别造成一定的影响。因此我们采用形态学滤波过滤超声图像周边无效的噪声信息，主要提取超声图像中间区域的有效信息。



图 12 超声图像随机截取前后的图像

Fig. 12 Ultrasound images before and after random sampling

形态学滤波可用于检测图像中指定的特征，抑制图像中的噪声信息等。膨胀、腐蚀、开闭运算是形态学中基本的变化形式，也有一些高级的转化，例如凸包、连通区域标记、删除小

块区域等。我们采用的基于形态学滤波截取图像流程如图13所示，首先对原始图像进行边缘增强，增强图像的对比度，然后利用形态学中的凸包求出图像的最大边界（联通区域），然后对图像进行截取，这是一种具有指导性的截取，而不是随机截取，图14为利用形态学滤波对某图像进行截取的过程。

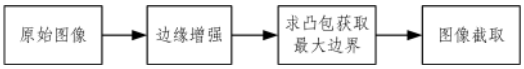


图13 基于形态学滤波截取图像流程图

Fig. 13 Flow chart of image capture based on morphological filtering

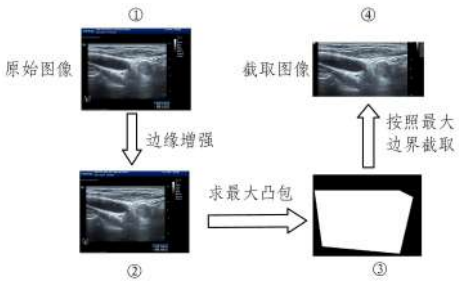


图14 基于形态学滤波截取图像示例

Fig. 14 Sample image capture based on morphological filtering

基于上面利用形态学滤波截取的方法，我们再次对数据集进行数据增强，实验结果如表4和图15所示。

表4 不同数据集划分比例下基于两种不同方式进行数据增强前后的 *mAP* 比较

Table 4 *mAP* comparison of different data set partition ratios based on two different ways of data enhancement

（单位：%）

是否数据增强	划分比例（训练集：验证集：测试集）			
	2.5∶2.5∶5	5∶2∶3	4∶4∶2	4∶3∶3
否	0.775	0.782	0.848	0.778
是（随机截取）	0.862	0.902	0.802	0.889
是（形态学滤波截取）	0.902	0.877	0.899	0.896

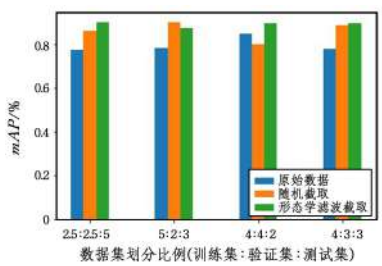


图15 不同数据集划分比例下基于不同截取方式的 *mAP*

Fig. 15 *mAP* histogram with different data set partition ratios based different truncation methods

从表4和图15可知，采用基于形态学滤波截取的数据增强对模型的准确率有一定的提升，4种划分比例的准确率基本接近90%，比基于随机截取的数据增强要好一些，而且更加稳定。基于形态学滤波的数据增强的4种划分比例的准确率基本相近，而基于随机截取的数据增强在4种划分比例下的准确率起伏较大，例如划分比例为4∶4∶2的准确率波动比较大。除此之外，我们对3种不同截取方式的4种划分比例的 *Loss* 收敛变化做了分析，如图16所示。可以看出，不同截取方式对于4种划分比例的收敛速度基本没影响，基本在3000次迭代时达到收敛。

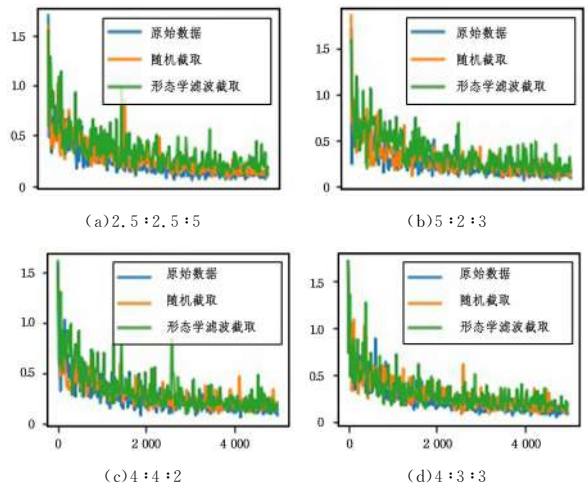


图16 不同截取方式对于划分比例的 *Loss* 变化折线图

Fig. 16 Polyline diagram of *Loss* variations for partition ratios for different truncation methods

（2）学习率和动量

除了上述的3种不同改进方法，一些超参数也会对模型的准确率有影响，例如学习率、动量、滑动窗口大小、平衡权重因子等。下面我们对学习率进行分析，一般来说，学习率越大，网络学习的速度越大。学习率直接影响模型能够以多快的速度收敛到局部最小值，如果学习率太大，有可能取不到极值，会在某一位置反复震荡，而学习率过小，则网络收敛速度太慢，还有可能陷入局部最优。选择合适的学习率不仅能加快收敛速度，减少网络训练时间，还会提高模型的准确率。

下面将数据集按照4∶3∶3的比例划分为训练集、验证集以及测试集，使用基于ZF网络的Faster R-CNN对基于形态学滤波截取的数据集分别采用0.1，0.005，0.001，0.0001的学习率进行5000次迭代训练，实验结果如表5和图17所示。

表5 不同学习率的 *mAP*

Table 5 *mAP* comparison of different learning rates

（单位：%）

网络模型	学习率			
	0.1	0.005	0.001	0.0001
ZF	None	0.902	0.889	0.760

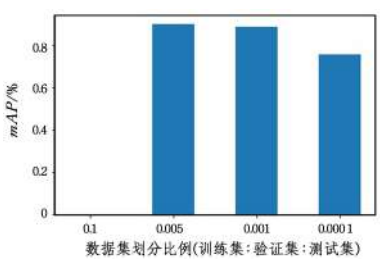


图17 不同学习率的 *mAP*

Fig. 17 *mAP* bars for different learning rates

从表5和图17发现，当学习率为0.1时结果为None，原因是学习率过大造成浮点数溢出。而学习率为0.005时准确率最高，而学习率为0.0001时准确率下降。我们对3种学习率的 *Loss* 收敛速度进行分析，如图18所示。可以看出，当学习率为0.0001时网络收敛速度非常慢，在迭代5000次时还未收敛完成，而学习率为0.001或者0.005时在迭代5000次左右时已经基本收敛达到稳定状态。

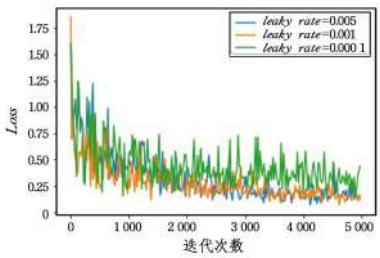


图 18 不同学习率的 *Loss* 变化折线图

Fig. 18 Polyline graphs of *Loss* changes for different learning rates

3.3 实验总结

上文对分别使用不同的卷积神经网络作为 Faster R-CNN 的特征提取部分，发现模型的结果不太理想，然后分别采用不同的划分比例，基于随机截取以及形态学滤波截取两种数据增强方法和不同的学习率对模型进行了改进，改进后的模型要比原模型更好，准确率更高，而且模型也相对更稳定。从实验中也可以发现：

(1)更深的网络有助于模型的学习，通过增加网络层数可以使得网络提取到更深层语义更丰富的信息，从而提高模型的准确率。

(2)当数据集较小时，不同的数据集划分比例会对模型的训练造成一定的影响，当划分比例过大时容易造成过拟合，当划分比例过小时容易造成欠拟合，都不利于提高模型的泛化能力。我们可以采取数据增强的方式增大数据集，通常的方式有水平翻转、垂直翻转、随机截取、固定截取、尺度变换、旋转、色彩抖动、边缘增强等。而随机截取的方式有可能使得截取到的信息不完整，可以采用基于形态学滤波的图像处理方式对图像进行截取，减少信息不完整的情况。通过数据增强往往能够提高模型的准确率，增强对图像形变的鲁棒性，提升模型的泛化能力，而基于形态学滤波的数据增强使得模型在提高准确率的同时更加稳定。

(3)一些超参数也会影响模型的效果，例如学习率会对模型的收敛速度产生影响。学习率过大时会造成网络产生振荡，取不到极值，还可能出现溢出，而学习率过小时网络收敛速度过慢，训练时间太长，而且还可能陷入局部最优，所以用合适的学习率对模型进行训练，才能取得更好的效果。

结束语 本文以甲状腺结节超声图像为主要研究对象，基于深度学习图像目标检测技术中的 Faster R-CNN 算法对甲状腺超声图像进行肿瘤定位以及基于 Caffenet 卷积神经网络对超声图像进行分类，采用深度学习框架 Caffe 以及 Django 框架构建肿瘤图像识别系统，用于识别甲状腺肿瘤超声图像。本文研究的重点主要有：

(1)分析传统的医疗图像识别技术，试图将深度学习目标检测技术应用于医疗图像领域。由于传统技术需要手工提取精确的特征，耗费精力大，而且精确度不高，利用深度学习中的卷积神经网络可以自学习图像特征的特点，通过增加网络层次获取更抽象和丰富的图像语义信息，不仅避免了数据预处理中的特征提取过程，还可以端到端地学习训练，不需要人工干预。

(2)基于 Faster R-CNN 的思想，采用 RPN 网络对超声图像产生用于 Fast R-CNN 判断的候选区域，提高了传统候选区域提取的时间，利用近似联合训练方法减少了网络收敛时间，并对不同的卷积网络、数据集划分形式、学习率等因素进行对比分析。实现发现，由于 CNN 底层是对图像的边缘信息和轮廓信息的提取，更高层是对图像语义信息和本质信息的抽象，所以利用深层 CNN 更能表征图像语义特征，准确率更好；由于数据集小，不同的数据集划分形式也会对模型产生一定的影响，采用数据增强技术之后可以增加样本的多样性，提高模型的准确率以及泛化能力。

(3)在数据增强中，随机截取的方式可能会造成图像信息不完整，无效的噪声信息也会影响模型的判断。基于传统处理技术，提出基于形态学滤波截取方法，利用形态学滤波中的凸包算法求出图像的最大凸包，根据凸包的最大边界截取图像，避免了图像信息截取不完整的情况，而且能够去除超声图像周边的无效噪声信息。

在以后的工作中，我们将继续研究该算法在 Map-Reduce 框架下的分布式训练和分类。该方法能够快速获得识别结果，提高了该研究的应用前景。此外，我们还可以将该技术框架进一步应用到超声视频识别中。

参 考 文 献

[1] ALISON N J, DJAMAL B. Ultrasound image segmentation: a survey[J]. IEEE Transactions on Medical Imaging, 2006, 25(8): 987-1010.

[2] YU-LEN H, DAR-REN C. Watershed segmentation for breast tumorin 2-d sonography[J]. Ultrasound in Medicine & Biology, 2004, 30(5): 625-632.

[3] GAETANO R, MASI G, POGGI G, et al. Markercontrolledwatershed-based segmentation of multiresolution remote sensingimages[J]. IEEE Transactions on Geoscience & Remote Sensing, 2015, 53(6): 2987-3004.

[4] NAIMI H, ADAMOU-MITICHE A B H, MITICHE L. Medical imagedenoising using Dual Tree Complex Thresholding Wavelet Transformand Wiener filter[J]. Journal of King Saud University-Computer and Information Sciences, 2015, 27(1): 40-45.

[5] PHAM V N, LONG T N, NGUYEN T D. Feature-reduction fuzzyco-clustering algorithm for hyperspectral image segmentation[C]// IEEE International Conference on Fuzzy Systems. 2017.

[6] ZHAO Y, RADA L, CHEN K, et al. Automatedvessel segmentation using infinite perimeter active contour model with hybrid region information with application to retinal images[J]. IEEE Transactions on Medical Imaging, 2015, 34(9): 1797-1807.

[7] FAN X, JU L, WANG X, et al. A fuzzy edge-weighted centroidalvoronoi tessellation model for image segmentation[J]. Computers & Mathematics with Applications, 2016, 71(11): 2272-2284.

[8] WANG H, HUANG T Z, XU Z, et al. An active contour modeland its algorithms with local and global gaussian distribution fitting energies[J]. Information Sciences An International Journal, 2014, 263(1): 43-59.

[9] KONUR U, GRGEN F S, VAROL F, et al. Computer aided detection of spina bifida using nearest neighbor classification withcurvature scale space features of fetal skulls extracted from ultrasound images[J]. Knowledge-Based Systems, 2015, 85(C): 80-95.

[10] FAROKHI S, SHEIKH U U, FLUSSER J, et al. Near infrared

face recognition using zernike moments and hermite kernels[J]. Information Sciences An International Journal, 2015, 316(C): 234-245.

[11] WENG T, YUAN Y, LING S, et al. Clothing image retrievalusing color moment[C]// International Conference on Computer Science & Network Technology. 2014.

[12] YAMAGUCHI J, YONEYAMA A, MINAMOTO T. Automatic detection of early esophageal cancer from endoscope image using fractal dimension and discrete wavelet transform[C]// International Conference on Information Technology-new Generations. 2015.

[13] DAS J, ROY H. Human face detection in color images using hsvcolor histogram and wld[C]// International Conference on Computational Intelligence & Communication Networks. 2015.

[14] XU X, QUAN C, REN F. Facial expression recognition based ongabor wavelet transform and histogram of oriented gradients [C]// IEEE International Conference on Mechatronics & Automation. 2015.

[15] EMMANUEL A, OLUGBARA O O. Lung cancer prediction usingneural network ensemble with histogram of oriented gradient genomicfeatures[J]. The Scientific World Journal, 2015, 2015: 786013.

[16] MOCANU D C, AMMAR H B, LOWET D, et al. Factored four way conditional restricted Boltzmann machines for activity recognition[J]. Pattern Recognition Letters, 2015, 66(C): 100-108.

[17] KOREZ R, LIKAR B, PERNU F, et al. Model-based segmentationof vertebral bodies from mr images with 3d cnns[C]// International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2016: 433-441.

[18] MOESKOPS P, WOLTERINK J M, VELDEN B H M V D, et al. Deep learningfor multi-task medical image segmentation in multiple modalities[C]// International Conference on Medical Image Computing & Computerassisted Intervention. 2016.

[19] GULSHAN V, PENG L, CORAM M, et al. Development and validation of a deep learning algorithm fordetection of diabetic retinopathy in retinal fundus photographs[J]. Jama, 2016, 316(22): 2402.

[20] ESTEVA A, KUPREL B, NOVOA R A, et al. Dermatologist-level classification of skin cancer with deep neural networks[J]. Nature, 2017, 542(7639): 115-118.

[21] MANSANET J, ALBIOL A, PAREDES R, et al. Mask selective regularization for restricted boltzmann machines[J]. Neurocomputing, 2015, 165(C): 375-383.

[22] ANTONY J, MCGUINNESS K, CONNOR N E O, et al. Quantifying radio graphic knee osteoarthritis severity using deep convolutional neural networks[J]. 2016.

[23] BENGTSSON E, MALM P. Screening for cervical cancer using automated analysis of pap-smears[J]. Computational and Mathematical Methodsin Medicine, 2014, 2014(2962): 842037.

[24] KOOI T, LITJENS G, GINNEKEN B V, et al. Large scale deep learning for computer aided detection of mammographic lesions [J]. Medical Image Analysis, 2017, 35: 303-312.



GU Wan-rong, born in 1982, Ph.D, assistant professor. His main research interests include machine learning, information retrieval and recommendation.



MAO Yi-jun, born in 1979, Ph.D, assistant professor. His main research interests include machine learning, bioinformatics and algorithm.

(上接第 249 页)

[6] MICHAILOVICH O, RATHI Y, TANNENBAUM A. Image Segmentation Using Active Contours Driven by The Bhattacharyya Gradient Flow[J]. IEEE Transactions on Image Processing, 2007, 16(11): 2787-2801.

[7] WU H, APPIA V, YEZZI A. Numerical Conditioning Problems and Solutions for Nonparametric i. i. d. Statistical Active Contours[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(6): 1298-1311.

[8] ZHOU Z, DAI M, WANG T, et al. Prior Distribution-Based Statistical Active Contour Model[J/OL]. Multimedia Tools and Application: 1-24. https://10.1007/S11042-019-08101-2.

[9] OSHER S, SETHIAN J. Fronts Propagating with Curvature Dependent Speed: Algorithms Based on Hamilton-Jacobi Formulations[J]. Journal of Computational Physics, 1998, 79 (1): 12-49.

[10] PARZEN E. On Estimation of a Probability Density Function and Mode[J]. The Annals of Mathematical Statistics, 1962, 33(3): 1065-1076.

[11] HUANG Y, QIU C, GUO Y, et al. Surface Defect Saliency of Magnetic Tile[J]. The Visual Computer, 2018: 1-12.

[12] KUMAR A, PANG G. Defect Detection in Textured Materials Using Gabor Filters[C]// Industry Applications Conference. IEEE, 2000: 1041-1047.

[13] CHENG M, MITRA N. Global Contrast Based Salient Region Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 37(3): 409-416.

[14] GONG M, LI H, ZHANG X, et al. Nonparametric Statistical Active Contour Based on Inclusion Degree of Fuzzy Sets[J]. IEEE Transactions on Fuzzy Systems, 2016, 24(5): 1176-1192.

[15] CHAN T, ZHU W. Level Set Based Shape Prior Segmentation [C]// Computer Vision and Pattern Recognition. IEEE, 2005, 2: 1164-1170.

[16] ZHAO H K, CHAN T, MERRIMAN B, OSHER. S, A Variational Level Set Approach to Multiphase Motion[J]. Journal of Computational Physics, 1996, 127: 179-195.



YANG Zhi-wei, born in 1968, bachelor, senior engineer. His main research interests include industrial internet, 5G applications and planning, consulting, design of smart cities, also with the establishment of industrialization and information integration system.

ISSN 1000-565X
CODEN HLDKEZ

華南理工大學學報

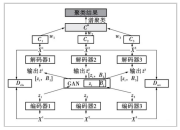
JOURNAL OF SOUTH CHINA UNIVERSITY OF TECHNOLOGY

自然科学版

Natural Science Edition

第48卷
1
2020年1月
Vol. 48 No. 1 (January 2020)
中国·广州 GUANGZHOU CHINA

基于 StarGAN 和子空间学习的缺失多视图聚类 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200128)


(https://zrb.bjb.scut.edu.cn/fileup/1000-565X/PIC/2020-0128_pic.jpg)

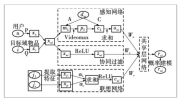
刘小兰 叶泽慧

2020, 48(11): 87-98. doi:10.12141/j.issn.1000-565X.200128 (https://doi.org/10.12141/j.issn.1000-565X.200128)

摘要 (319) PDF (743KB) (65)

相关文章 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200128#RelatedCitationTab) | 多维度评价 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200128#MetricsTab)

推荐系统信息跨领域的改进迁移学习算法 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.190619)


(https://zrb.bjb.scut.edu.cn/fileup/1000-565X/PIC/2019-0619_pic.jpg)

张子烨, 李明畅, 梁凌睿, 等

2020, 48(11): 99-106. doi:10.12141/j.issn.1000-565X.190619 (https://doi.org/10.12141/j.issn.1000-565X.190619)

摘要 (465) PDF (562KB) (253)

相关文章 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.190619#RelatedCitationTab) | 多维度评价 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.190619#MetricsTab)

基于时间卷积神经网络的短时交通流预测算法 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.190276)


(https://zrb.bjb.scut.edu.cn/fileup/1000-565X/PIC/2019-0276_pic.jpg)

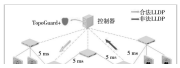
袁华 陈泽濠

2020, 48(11): 107-113,122. doi:10.12141/j.issn.1000-565X.190276 (https://doi.org/10.12141/j.issn.1000-565X.190276)

摘要 (482) PDF (618KB) (171)

相关文章 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.190276#RelatedCitationTab) | 多维度评价 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.190276#MetricsTab)

SDN 拓扑攻击及其防御 (https://zrb.bjb.scut.edu.cn/CN/10.12141/j.issn.1000-565X.200282)



陆以勤, 毛中书, 程喆, 等

2020, 48(11): 114-122. doi:10.12141/j.issn.1000-565X.200282 (https://doi.org/10.12141/j.issn.1000-565X.200282)

通知公告

本刊2023年第10期被EI收录论文目录 (https://zrb.bjb.scut.edu.cn/CN/news 2023-12-28

本刊2023年第9期被EI收录论文目录 (https://zrb.bjb.scut.edu.cn/CN/news 2023-12-01

本刊2023年第8期被EI收录论文目录 (https://zrb.bjb.scut.edu.cn/CN/news 2023-10-31

本刊2023年第7期被EI收录论文目录 (https://zrb.bjb.scut.edu.cn/CN/news

更多... (../news/folde

精彩推荐

车路协同系统关键技术研究进展 (../10.12141/j.issn.1000-565X.2302

考虑道路负载均衡的码头多AGV无冲 (../10.12141/j.issn.1000-565X.2302

TSN中基于链路负载均衡的AVB流量 (../10.12141/j.issn.1000-565X.2207

一种面向中文拼写纠错的自监督预训 (../10.12141/j.issn.1000-565X.2300

多类型公交线路的运行可靠性优化策 (../10.12141/j.issn.1000-565X.2204

面向边缘特征的实时模板匹配方法 (../10.12141/j.issn.1000-565X.2207

更多>> (../article/sho

阅读排行

第 48 卷 第 11 期
2020 年 11 月

华南理工大学学报(自然科学版)
Journal of South China University of Technology
(Natural Science Edition)

Vol. 48 No. 11
November 2020

doi: 10.12141/j.issn.1000-565X.190619

推荐系统信息跨领域的改进迁移学习算法

张子烨[1] 李明畅[1] 梁凌睿[1] 张铭华[1] 谢贤芬[2] 古万荣[3†]

(1. 华南理工大学 数学学院，广东 广州 510640；2. 暨南大学 经济学院，广东 广州 510632；
3. 华南农业大学 数学与信息学院，广东 广州 510642)

摘 要：推荐技术中的单领域推荐算法面临诸多现实问题，主流的协同过滤算法只利用了用户与物品的交互的信息，无法避免会受到实际应用场景中数据稀疏性的影响。本文致力于融合其他领域的信息来处理稀疏数据，解决了现有算法较难从跨领域中找到有效关联的问题。文中提出了一种联想型感知网络模型，它基于深度神经网络，通过深度挖掘物品的内容信息和信息跨领域的方法来实现迁移学习，获取最优化的特征关联，进而优化了推荐的准确性。该算法很好地处理了推荐系统中数据稀疏性问题，并且在公测数据集的实验中比多种新近算法提升达到 15% ~20%，表现出了更好的性能和可扩展性。

关键词：推荐系统；数据挖掘；机器学习；神经网络

中图分类号：TP391 **文章编号**：1000-565X(2020)011-0099-08

推荐系统广泛地应用于电子商务、视频网站、音乐电台、广告投放、新闻咨询等领域，为处于这些领域的公司的盈利做出了可观的贡献。据亚马逊前首席科学家 Andreas Weigend 的说法，亚马逊每年有 20% ~30% 的销售来自推荐系统（参见 blog.Kiwitobes.com/?p=58）。这足以说明对推荐系统进行研究能带来经济上的效益。

协同过滤算法通过用户与物品交互的行为预测用户的偏好，此法与物品内容无关。神经网络的引入，将向量（用于储存用户或物品特征）的学习程度发展到了非线性表示[1-2]，其中的基本体系结构被扩展到了利用社会和地理信息来提取用户和物品的信息[3]。一些神经网络算法主要学习基于显性反馈的评分预测任务[4-5]。文中的内容专注于从隐性反馈中学习 Top-N 列表推荐[6]。然而，协同过滤算法存在着数据稀疏的问题。

物品通常与内容信息（如非结构化的文本）相关联，如论文关键词和商品广告文案。协同过滤算法可以扩展利用物品内容信息[7-9]和用户评论[10-12]。Chelliah 等[13]对当前基于用户评论增强商品推荐的研究做了一个总结。

记忆网络可以用领域外的感知信息进行推理。由于记忆网络具有学习词语语义的能力，可以解决词的稀疏性和同义词、近义词的问题，因此可以用记忆网络对物品内容或相似用户进行建模[14-15]。文中沿着这一研究思路，利用神经网络从文本内容中提取重要信息。

随着神经网络[16]技术的兴起，许多神经网络的文本挖掘技术被应用到推荐系统中来。在基于文本处理的推荐系统，CNN[17-19]文本处理网络被用来提升单领域推荐中评分预测的精度。

跨领域推荐是缓解数据稀疏性的有效方案，其

收稿日期：2019-09-18

基金项目：广东省科技计划项目（2018A070712021）

Foundation item: Supported by the Science and Technology Planning Project of Guangdong Province (2018A070712021)

作者简介：张子烨（1998-），博士生，主要从事推荐系统、生存分析领域研究。E-mail: ma_zere_x1@mail.scut.edu.cn

† 通信作者：古万荣（1982-），博士，讲师，主要从事大数据分析、信息检索和推荐系统研究。E-mail: guwanrong@scau.edu.cn

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

中一种有效的方法是将目标域上的协同过滤算法和辅助域上的另一个协同过滤算法组合在一起，从而实现了知识的迁移[20]。

在跨领域推荐系统，给定两个不同任务的处理方案，就可以让跨缝卷积网络(CSN)[21-22]尝试对这两个方案进行不同权重的线性组合，并将这些组合作为输入信息提供给连接两个领域的神经网络，从而实现两个领域之间的知识迁移。近期，Khan等[23]对跨领域推荐系统的研究进行了系统的归纳。沿着这个思路，文中从提取项目的内容信息和信息跨领域的这两个方面对推荐系统进行研究。

为此，文中提出了一种改进的神经网络模型:联想型感知网络模型，它用于跨领域的内容推荐，其中感知网络是一种记忆网络，它能够有效地提取物品的内容信息; 而联想网络起到知识迁移的作用，它能跨领域传输选定的信息。它的前部由共享的特性交互层构成，连接了从每个网络中学习到的成果。在真实的数据集上，文中的联想型感知网络与各种其它算法模型进行比较，在多个指标方面展现出了更好的性能。

1 建模

文中将在本节中描述所提出的联想型感知网络模型，该模型利用了物品的内容信息。

表1提及的参数均可由实验调取的数据集获取，具体生成方式在下文提及。在介绍网络体系结构之前，先描述推荐系统问题和模型公式。

表1 模型参数

Table 1 Model parameters

参数	维度	参数说明
$\boldsymbol{P}$	$m \times d$	用户嵌入矩阵
$\boldsymbol{Q}$	$n_{\mathrm{T}} \times d$	目标域商品嵌入矩阵
$\boldsymbol{H}$	$n_{\mathrm{S}} \times d$	辅助域商品嵌入矩阵
$\boldsymbol{A}$	$L \times 2d$	内部存储器矩阵
$\boldsymbol{C}$	$L \times 2d$	外部存储器矩阵
$\boldsymbol{W}, b$	$2d \times 2d, d$	用户-物品交互的线性映射权重和误差
$\boldsymbol{W}_o, \boldsymbol{W}_z, \boldsymbol{W}_c$	$d \times d$	网络输出的线性映射
$\boldsymbol{h}$	$3d$	共享层权重

1.1 问题的提出

假设现在有一个目标域(如京东商品域)，用户-物品交互矩阵 $\boldsymbol{R}_{\mathrm{T}} \in \mathbf{R}^{m \times n_{\mathrm{T}}}$ 和一个辅助域(如腾讯视频域)的矩阵 $\boldsymbol{R}_{\mathrm{S}} \in \mathbf{R}^{m \times n_{\mathrm{S}}}$，其中 $m = |\mathcal{U}|$ 是用户集 $\mathcal{U}$ 的用户数量，$n_{\mathrm{T}} = |\mathcal{T}_{\mathrm{T}}|$ 是目标域物品 $\mathcal{T}_{\mathrm{T}}$ 的数量，$n_{\mathrm{S}} = |\mathcal{T}_{\mathrm{S}}|$ 是辅助域物品 $\mathcal{T}_{\mathrm{S}}$ 的数量。注意这里的用户指的是同时在京东和腾讯有账户的用户，所以这里只有一个用户集，这是文中进行知识迁移的前提。

现在使用 u 对用户集 $\mathcal{U}$ 进行检索，使用 i 对目标域的物品 $\mathcal{T}_{\mathrm{T}}$ 进行检索，使用 j 对辅助域的物品 $\mathcal{T}_{\mathrm{S}}$ 进行检索。

对于目标域而言，如果用户 u 对项目 i 产生了行为，则 $r_{ui} \in \{0, 1\}$ 的值为1，否则为0。设 $\mathcal{J}^u = (j_1, j_2, \cdots, j_s)$ 为被用户产生过行为的辅助域物品集。现在假定目标域除了有行为信息，还有内容信息，比如商品的广告文案、产品规格或商品评价，这里主要用到商品评价。用 d_{ui} 表示用户 u 对物品 i 产生行为的文本内容(假设语言是英文)，里面包含了一个单词序列 $d_{ui} = (w_1, w_2, \cdots, w_l)$，其中的每一个单词都来自一个词汇表 $\mathcal{V}$，设 $l = |d_{ui}|$ 是这段文本内容的长度。

对于推荐任务，文中的简单目的是对用户的历史行为加以分析，得出一个推荐列表推荐给用户(也就是Top-N推荐)。期望刚才从辅助域中得到的行为信息能提高算法在目标域上的推荐性能。

用向量 $\boldsymbol{r}_u = (r_{u1}, r_{u1}, \cdots, r_{un_{\mathrm{T}}})$ 表示用户 u 的交互行为，文中算法将以用户、内容文本、产生行为的辅助域物品作为条件，建立起观察该用户每次操作的概率模型。

$$\hat{r}_{ui} \triangleq p(r_{ui} = 1 \mid u, d_{ui}, \mathcal{J}^u) \tag{1}$$

这个公式能强化联想型感知网络模型的“想象力”，因为它会通过3个主要因素考察用户 u 喜欢目标域物品 i 的数学概率:

(1) 用户 u 本人的偏好(来自目标域)。

(2) 用户 u 对目标域物品(已购商品)的评价文本信息(d_{ui})。

(3) 用户 u 在辅助域的行为($\mathcal{J}^u$)。

文中将这个矩阵 $\boldsymbol{R}_{\mathrm{T}}$ 的似然函数定义为

$$p(\boldsymbol{R}_{\mathrm{T}}) = \prod_u \prod_i p(r_{ui} \mid u, d_{ui}, \mathcal{J}^u). \tag{2}$$

文中所提出的联想型感知网络是一种神经网络，用端到端方式学习条件概率:

$$\hat{r}_{ui} = f(u, i, d_{ui}, \mathcal{J}^u \mid \Theta_f) \tag{3}$$

总体来说整个联想型感知网络由感知网络 $\boldsymbol{o}_{ui} = f_{\mathrm{M}}(u, i, d_{ui} \mid \Theta_{\mathrm{M}})$($f_{\mathrm{M}}$ 表示数据经过感知网络处理)和联想网络 $\boldsymbol{c}_{ui} = f_{\mathrm{T}}(i, \mathcal{J}^u \mid \Theta_{\mathrm{T}})$($f_{\mathrm{T}}$ 表示数据经过联想网络处理)组成。一个共享层 $f_{\mathrm{S}}(\boldsymbol{o}_{ui}, \boldsymbol{z}_{ui}, \boldsymbol{c}_{ui} \mid \Theta_s)$($f_{\mathrm{S}}$ 表示数据经过共享层网络的处理)共

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

享了交互特性，领导着每一个个体的网络学习，其中 z_{ui} 是 (u, i) 交互行为的非线性表示。

1.2 感知网络的构建

文中引入感知网络来处理上述从目标域中提取出来的非结构化文本。感知网络是一种带有记忆增强的神经网络变体，它可以学习用户 - 物品交互信息（非结构化文本），并得出抽象表示。感知网络中的的注意力机制可以自动识别哪些词与用户偏好密切相关。文中构建感知网络的想法与 Hu 等[15]的想法不谋而合，但角度有所不同。

感知网络使用一个矩阵 $\boldsymbol{A} \in \mathbf{R}^{L \times 2d}$ 作为存储器，其中 L 为词汇量，$2d$ 为矩阵的维数。感知网络还有一个外部存储器 C，其维数与 $\boldsymbol{A}$ 相同。

现给定一个文件 $d_{ui} = (w_1, w_2, \cdots, w_l)$ 来储存 (u, i) 的交互信息，文中建立了一个感知存储单元 $m_k \in \mathbf{R}^{2d}$，其中每个单词 w_k 都会被映射到矩阵 $\boldsymbol{A}$ 的向量中，而 $k = 1, 2, \cdots, l$ 和最长的文件的长度是感知的最大容量 $l_{\max}$。文中建立了一个偏好向量 $\boldsymbol{q}$ 用于对应文件 d_{ui} 和用户 - 物品交互行为 (u, i)，其中每个元素 $\boldsymbol{q}_k$ 记录了用户 u 与物品 i 的相关性，计算方法如下:

$$\boldsymbol{q}_k = \boldsymbol{x}_u^{\mathrm{T}} \boldsymbol{m}_k^{(u)} + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{m}_k^{(i)} \tag{4}$$

式中，$\boldsymbol{m}_k = [\boldsymbol{m}_k^{(u)}, \boldsymbol{m}_k^{(i)}]$ 包含用户部分 $\boldsymbol{m}_k^{(u)}$ 和物品部分 $\boldsymbol{m}_k^{(i)}$。用户和物品分别以 $\boldsymbol{x}_u$ 和 $\boldsymbol{x}_i$ 嵌入矩阵 $\boldsymbol{P} \in \mathbf{R}^{m \times d}$ 和矩阵 $\boldsymbol{Q} \in \mathbf{R}^{n_{\mathrm{T}} \times d}$ 中。在式（4）的右端，第 1 项匹配了用户 u 和单词语义。例如，用户是一位足球球迷，他可能会对“球鞋”、“足球”这样的词感兴趣，而不是“高跟鞋”、“口红”。第 2 项计算物品 i 与单词语义的相关度。例如这个商品是一款球鞋，那么与之相关的词应该是“尺码”、“弹性”之类的，而不是“色号”、“保湿”。

然后可以通过一个向量点积来压缩上面的那两项，此举将用户和物品的关系压缩到了 $\boldsymbol{x}_{ui} = [\boldsymbol{x}_u, \boldsymbol{x}_i]$ 中:

$$\boldsymbol{q}_k = \boldsymbol{x}_{ui}^{\mathrm{T}} \boldsymbol{m}_k, \quad k = 1, 2, \cdots, l \tag{5}$$

注意力机制可以学习词义的权重，然后关注其中重要的词语。传统的词汇集合预先定义了一个初始权重函数，如平均值函数。而文中通过注意力机制产生的权重来推断每个单词的重要性:

$$p_k = \mathrm{Videomax}(q_k) = \frac{\exp(\beta q_k)}{\sum_{k'=1}^{l} \exp(\beta q_{k'})} \tag{6}$$

先在 d_{ui} 中的单词上生成一个概率分布。注意力机制可以使感知网络重点关注某一些词汇。当归一化指数函数的指数特别大的时候，可通过引入参数 β 来稳定数值计算，它可以放大或者减弱注意力，比如商品销量太高或者太低都会触发这一机制，使注意力的概率分布变弱。在这里文中根据维数来按比例设置参数 $\beta = d^{-\frac{1}{2}}$ [24]。

文中用注意力机制对外部存储器进行了插值:

$$o_{ui} = \sum_k p_k \boldsymbol{c}_k \tag{7}$$

式中，外部存储器单元 $\boldsymbol{c}_k \in \mathbf{R}^d$ 是词向量 $\boldsymbol{w}_k$ 另外的嵌入向量，由矩阵 $\boldsymbol{C}$ 的映射形成。此举生成了一个抽象的表示作为感知网络的输出内容。

外部存储器允许长时间储存知识信息，特别是每个词汇在匹配用户偏好时的重要性信息。也就是说，通过注意力机制遴选出来的都是重要的词汇，值得被收入矩阵 $\boldsymbol{C}$ 中妥善保存。

注意力机制对不同的用户和不同的物品可以对不同的词语进行加权，最终输出的 o_{ui} 是一种高级的信息，这些信息正是从用户 - 物品交互行为 (u, i) 中的文本单词 d_{ui} 中凝炼出来的。

1.3 联想网络的构建

联想网络的主要作用是获取辅助域的信息并加以利用，它是一种有选择性的跨领域传输知识信息的神经网络。它的主要构建想法是在训练过程中不断选择辅助域中用户 - 物品交互信息的权重并加以调整。假设用户 u 在辅助域中与物品 $\boldsymbol{j}^u = (j_1, j_2, \cdots, j_s)$ 进行了交互，联想网络则会通过迁移向量 $\boldsymbol{c}_{ui} \in \mathbf{R}^d$ 来判定目标域中物品 i 是否会被对辅助域中某物品 j 产生过行为的用户 u 喜欢上。例如，当文中想要预测一个用户是否会对京东中的某款球鞋感兴趣时，文中会考察他在腾讯视频看了哪位球员的精彩集锦，这比考虑他在腾讯视频看过哪些电影重要得多。目标域和辅助域物品之间的相似度可以通过以下点积算出来:

$$a_j^{(i)} = x_i^{\mathrm{T}} x_j, \quad j = 1, 2, \cdots, s \tag{8}$$

式中，$x_j \in \mathbf{R}^d$ 将辅助域物品 j 嵌入矩阵 $\boldsymbol{H} \in \mathbf{R}^{n_{\mathrm{S}} \times d}$ 中。该公式计算了目标域物品与用户产生过行为的辅助域项目之间的匹配度。例如目标商品 C 罗联名球鞋的相似度与辅助域视频《C 罗在皇马的十佳进球》的关系大一些（对用户 u 来说），而不是《C 罗代言某洗发水》这种视频。将相似性得分归一化为辅助域物品的概率分布:

$$\alpha_j = \mathrm{Videomax}(\alpha_j^{(i)}) \tag{9}$$

然后迁移向量就能用相应辅助域物品的加权再求和算出来:

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

$$\boldsymbol{c}_{ui} = \mathrm{ReLU}\left(\sum_j \alpha_j \boldsymbol{x}_j\right) \tag{10}$$

文中用激活函数整流线性单元引入了非线性迁移向量。根据以往的情况，激活函数在 $\mathrm{ReLU}(x) = \max(0, x)$ 时效果比较好，因为它的非饱和性质适合于稀疏的数据。

迁移向量 $\boldsymbol{c}_{ui}$ 是一种高级表达，它总结了辅助域的知识，并作为联想网络的输出项。联想网络可以在目标域用户 - 物品的交互信息中，有选择地传输辅助域中的知识信息。

1.4 联想型感知网络的构建

总的来说这个网络是一个多层次前馈神经网络，它的大致结构如图 1 所示。输入层是用户 u、目标域物品 i 和对应的辅助域物品集合 $[j]^u = (j_1, j_2, \cdots, j_s)$。内容文本组 d_{ui} 是感知网络生成的抽象表示 $\boldsymbol{o}_{ui}$。然后在联想网络组 (u, i) 的指导下，将辅助域物品信息写入迁移向量 $\boldsymbol{c}_{ui}$ 中。

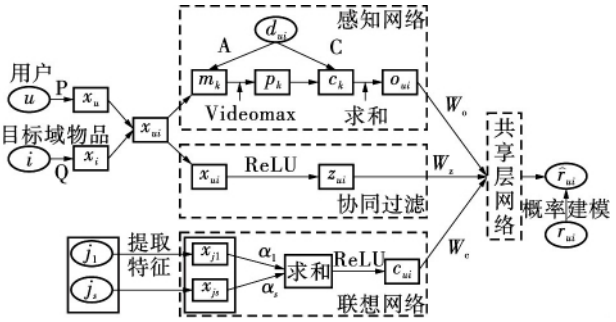


图 1 网络示意图

Fig. 1 Schematic diagram of the network

为了详解联想型感知网络，先用一层隐藏层网络去学习一个用户 - 物品的交互的非线性表示:

$$\boldsymbol{z}_{ui} = \mathrm{ReLU}(\boldsymbol{W}\boldsymbol{x}_{ui} + \boldsymbol{b}) \tag{11}$$

式中，$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 是隐藏层中的权重和误差参数。在典型的塔式结构中，$\boldsymbol{z}_{ui}$ 的长度一般是 $\boldsymbol{x}_{ui}$ 的一半。

这 3 个独立的网络的输出成果是内容文本、辅助域知识信息和用户 - 物品交互信息的高级特性，是不同网络学习到的不同特征空间，因此需要用一个共享层联系它们。

$$\hat{r}_{ui} = \frac{1}{1 + \exp(-\boldsymbol{h}^{\mathrm{T}}\boldsymbol{y}_{ui})} \tag{12}$$

式中，$\boldsymbol{h}$ 是学习过程中自主调整的参数。联立之后文中可以得到:

$$\boldsymbol{y}_{ui} = [\boldsymbol{W}_o\boldsymbol{o}_{ui}, \boldsymbol{W}_z\boldsymbol{z}_{ui}, \boldsymbol{W}_c\boldsymbol{c}_{ui}] \tag{13}$$

1.5 模型的学习

考虑到本次项目推荐任务利用的是隐性反馈，用平方损失 $(\hat{r}_{ui} - r_{ui})^2$ 来衡量可能有失公正。因为它一般是用在商品等物品的打分预测上面的。所以，文中这次主要使用二元交叉熵损失:

$$\mathcal{L} = -\sum_{(u,i)\in S} r_{ui}\log\hat{r}_{ui} + (1 - r_{ui})\log(1 - \hat{r}_{ui}) \tag{14}$$

其中训练样本 $S = \boldsymbol{R}_{\mathrm{T}}^{+} \cup \boldsymbol{R}_{\mathrm{T}}^{-}$ 是观测目标的交互矩阵和随机采样选出来的负样本的结合。一般 $|\boldsymbol{R}_{\mathrm{T}}^{+}| = |\boldsymbol{R}_{\mathrm{T}}^{-}|$，且文中不会事先预定义负样本，因为这只能生成一个固定的负样本训练集。可以在每个单元中产生负样本，使不同的训练集的负样本都能被用上，避免浪费。

该目标函数有概率解释，是下列似然函数的负对数似然度:

$$L(\Theta \mid S) = \prod_{(u,i)\in\boldsymbol{R}_{\mathrm{T}}^{+}} \hat{r}_{ui} \prod_{(u,i)\in\boldsymbol{R}_{\mathrm{T}}^{-}} (1 - \hat{r}_{ui}) \tag{15}$$

其中模型参数是

$$\Theta = \{\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{H}, \boldsymbol{A}, \boldsymbol{C}, \boldsymbol{W}, \boldsymbol{b}, \boldsymbol{W}_o, \boldsymbol{W}_z, \boldsymbol{W}_c, \boldsymbol{h}\} \tag{16}$$

与式 (2) 相比，不使用空白信息去构造目标矩阵 $\boldsymbol{R}_{\mathrm{T}}$，而是仅从这些未观测样本的一个小子集中学习，并在每次优化迭代时从中随机抽取作为负样本 (即负采样技术)。

目标函数可以通过自适应矩随机梯度下降 (SGD) 及其变体 (如 Adam) 进行优化:

$$\Theta^{\mathrm{new}} \leftarrow \Theta^{\mathrm{old}} - \eta\frac{\partial L(\Theta)}{\partial\Theta} \tag{17}$$

式中，η 是学习率。调用的库会自动进行微分，故文中省略了使用梯度方程 $\frac{\partial L(\Theta)}{\partial\Theta}$ 处理的过程。

2 实验

这一部分将用实验考察联想型感知网络与最先进的推荐系统相比各项评估情况如何，以作为各推荐系统性能参考。本节先介绍评估标准和实验设置，然后再比较不同推荐系统的性能。

2.1 数据集及预处理

如表 2 所示，文中在 3 类数据集上进行实验。第 1 类数据集来自 CiteULike (详见 http://www.cs.cmu.edu/~chongw/data/citeulike/)，被广泛用于评估论文领域的推荐系统性能[8]。第 2 类数据集是开放的亚马逊数据，该数据集已经被广泛用于评估推荐系统的性能[10]，本次使用的是其中的男士分区和体育用品分区。数据集中不包含其他信息，如停留时间和人口统计学数据等上下文信息。第 3 类是由 Cheetah Mobile (美国一家大型互联网公

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

司）提供的移动网络数据集，里面包含了用户阅读新闻的记录和手机应用的安装历史（详见 http://www.cmcm.com/en-us/）。

表 2　数据集

Table 2　Datasets

领域	数据集	数据类型	数量
目标域	CiteULike	用户	5551
		标题	16980
		反馈	204986
		词数	1587000
	亚马逊男士	共同用户	8514
		商品（衣物）	28262
		评价（含评分）	56050
		词数	1845387
	手机新闻	共同用户	15890
		新闻	84802
		标题单词	612839
辅助域	亚马逊体育	共同用户	8514
		商品（用品）	41317
		评价（含评分）	81924
	手机应用	共同用户	15890
		应用	14340
		安装信息	817120

表 2 中总结了使用的数据集的大致情况。注意，3 类数据集的稀疏度都大于 99%，能充分考验算法的性能。

对于 CiteULike，文中使用 Wang 等[8]发布的版本，词汇量为 8 000 个，文本长度约为 160 万词，每单位文本（论文摘要）大概 90 词左右。注意这个数据集是单领域，文中用它来单独测试感知网络的性能。

对于亚马逊和 Cheetah Mobile 的数据集，文中仅使用其某个分区的数据，详见表 2。实验中过滤了停止词、语气词等无实义的单词，并使用 TF-IDF 方法选择前 8 000 个不同的单词作为词汇表。每个商品涉及的文本大概 33 词左右，每条新闻的标题不足 10 词。

第 1 类数据只被用在了拓展实验中。而在对比实验中，其他非跨域类算法将只使用亚马逊男士领域和手机新闻领域的数据。

2.2　评估方案

项目推荐任务采用领域内广泛采用的留一法（LOO）来进行评估，且遵循推荐系统中广泛接受的协议[25]。为此，在实验中为每个用户保留了他的一个交互动作（通常是最新的交互动作，如果没有时间信息，就采用随机选择）作为测试项。

按照标准的推荐策略，随机抽取 99 个不受用户影响的负面信息，然后评估推荐系统对这些负面信息的排名情况。仅选取推荐列表的前十个物品进行评估（即 TopN = 10）。

采用的评估项目是召回率（用 Hit Ratio 衡量）和归一化后的折算累计收益（NDCG）。前者召回率限于篇幅不予以介绍；后者取自折算累计收益（DCG），DCG 在不同的列表之间很难进行横向的评估。而文中评估一个推荐系统，不可能仅使用一个用户的推荐列表进行评估，所以不同用户的推荐列表的评估分数就需要进行归一化。

评估项目的相关代码都可以在 GitHub 上找到（详见 https://github.com/princewen/tensorflow_practice/tree/master/recommendation/Basic-Evaluation-metrics）。

2.3　参测算法介绍和准备

（1）最流行物品推荐（ItemPOP）。选取最热门的 10 个物品推荐给用户，热门程度通过该物品与所有用户的交互次数算出。最流行物品推荐算法较为简单，常用于各类算法的实验对比。

（2）多层感知方法（MLP）。它是一种使用前馈神经网络学习非线性相互作用的函数[25]。使用 GitHub 上发布的代码来评估（详见 https://github.com/hexiangnan/neural_collaborative_filtering）。此外，基本的 MLP 网络的隐藏层被设置为［64→32→16→8］，可从 MLP 相关文献中得出[28]。

（3）协同主题回归法（CTR）。这种模型算法利用一个辅助域来提高算法性能，CiteULike 数据集实际上来自此论文[7]。文中使用 GitHub 上发布的代码来评估（详见 http://www.cs.cmu.edu/~chongw/citeulike/）。

（4）本地集中式记忆推荐系统（LCMR）。该算法是一种混合算法，其本地网络类似于文中的感知网络，但是他们研究的重点是深度混合方法[15]。该算法是文中算法的主要竞争对手。该算法的设置方法基本与文中算法相似，亦可见于其原文的实验设置[15]。

文中采用相同的实验环境进行实验评估，即用运行在同一个 GPU 万丽 RTX2080 上的 Tensor Flow（详见 https://www.tensorflow.org）来评估。测试的参数由高斯随机数 $\mathcal{N}(0,0.01^2)$ 生成。优化器选择了 Adam[27]，学习率是 0.001，最小处理单元为 128 个，负采样率为 1，在验证集上的参数是 Φ =

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

$\{L,N,d\}$，各运行30遍取最高分。

2.4 评估结果及分析

在亚马逊和Cheetah Mobile数据集上的评估结果统计信息分别见于表3和表4。通过评估结果，文中可以得到以下初步结论:

(1) 文中的算法在亚马逊数据集上优于所有的参测算法，证明了文中算法在联系紧密的领域上的有效性。

(2) 亚马逊数据集是文中算法的优势领域，这是因为男士衣物和体育用品的关系相当紧密，十分有利于跨领域的算法。在该数据集上，文中的算法获得了约15% ~20%的提升。这说明了联想网络能够从辅助域中选择有用信息形成抽象表示，并且在选择时聚焦了与用户偏好强烈相关的重要词汇，从而能更有效地猜测用户的偏好。

(3) 文中方法与LCMR在移动网络数据集上差距很小。文中分析后认为原因有以下两点，第一是新闻领域中的用户大多立场明确，这有利于LCMR对用户的偏好进行提取，因为它专注识别文本信息; 第二是由于新闻领域的特殊性，在该领域中，所有的算法的成绩都会比较高，难以拉开差距，就例如最流行物品推荐在该领域中都能取得不错的效果。

表3 评估结果(亚马逊数据集)

Table 3 Test results (Amazon)

标准	ItemPOP	MLP	CTR	LCMR	本文方法	提升
召回率/%	20.06	28.11	28.15	28.29	34.79	22.98
NDCG/%	11.63	16.89	16.74	16.65	19.61	16.10

表4 评估结果(移动网络数据集)

Table 4 Test results (Mobile)

标准	ItemPOP	MLP	CTR	LCMR	本文方法	提升
召回率/%	45.54	61.61	62.33	63.09	64.45	2.16
NDCG/%	37.89	43.77	43.90	44.55	46.88	5.23

实验结果初步表明，文中提出的联想型感知网络在同时利用文本内容和辅助域知识进行推荐时具有优越性。并且，本算法适合在电子商务等用户偏好难提取，但领域间联系紧密的情形使用。

2.5 拓展实验

在本小节中，文中延续了2、3节的评估方案与评估设置，额外使用第1、2类数据集单独测试了联想型感知网络的两个重要构成部分: 感知网络和联想网络的性能。文中还测试了联想型感知网络在不同推荐列表长度下的性能表现，文中选取了TopN=5，10，15，20，并在第2类数据集上进行实验。文中只选取了最流行物品推荐和LCMR这两种算法参与本次拓展实验。

单独使用感知网络，即在不借助辅助域信息帮助的条件下，在目标域通过神经注意力机制匹配词义与用户偏好来完成推荐任务; 单独使用联想网络，即在不使用目标域用户-物品交互行为的前提下，通过分析用户在辅助域上的行为来完成推荐任务。其中联想网络和完整的联想型感知网络仅可用于含有辅助域的亚马逊数据集。

在CiteULike和亚马逊数据集上的实验结果见于表5和表6。通过评估结果，文中可以得到以下结论:

在第1类数据集上单独使用感知网络效果不佳，不及LCMR算法。文中分析后发现CiteULike数据集包含了很多文本信息(单位文本长度长达90词)，这十分有利于LCMR这种专注识别文本信息的混合推荐算法。而在亚马逊数据集上，文本信息相对减少，于是感知网络的性能得以反超LCMR。

表5 在论文数据集上的评估结果

Table 5 Test results on CiteULike

标准	ItemPOP	LCMR	感知网络
召回率/%	27.29	84.48	83.41
NDCG/%	18.28	60.96	59.87

表6 在亚马逊数据集上的评估结果

Table 6 Test results on Amazon

标准	ItemPOP	LCMR	感知网络	联想网络	本文方法
召回率/%	20.06	28.29	32.13	29.67	34.79
NDCG/%	11.63	16.65	19.01	17.95	19.61

不难看出在第2类数据集上，感知网络对文中算法的贡献比联想网络大。并且，在亚马逊男士和亚马逊体育这种联系紧密的领域中，感知网络、联想网络在单独使用时的性能均超过了LCMR，进一步说明了文中算法的优越性。

文中还改变了推荐列表的长度进行多次试验，得出了联想型感知网络不同的性能供参考。具体实验结果见于图2，实验结果表明，联想型感知网络可能在重视新颖性和多样性的任务中表现出较好的性能。这是因为提高新颖性和多样性要求推荐列表的长度要达到正常推荐列表长度(一般在10以内)的数倍。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

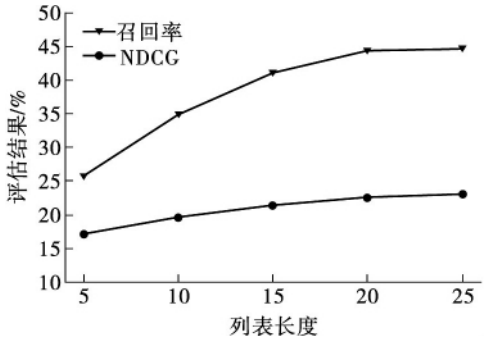


图 2　在不同列表长度下的评估结果

Fig. 2　Test results of different list lengths

3　结论

针对推荐算法的跨域推荐面临的数据稀疏性问题以及关联有效性问题，文中提出了一种新的深层次的神经网络模型：联想型感知网络模型，它由一个感知网络和一个联想网络组成，前者可以理解词语语义以匹配用户的偏好，后者可以有选择地在辅助域中传输有用信息帮助目标域。

实验结果表明，物品的内容信息和辅助领域的知识可以提高推荐系统的性能，并且可以在神经网络的体系结构下建起。利用这两种信息可以重置稀疏的用户－物品交互矩阵，解决了数据稀疏的问题。

文中的算法适合在电子商务等单个文本信息较少，但是领域之间联系紧密的情况下使用。在这样的领域，文中的算法表现出了更好的性能，提升达 15% ～20%。

此外，文中的算法在引入新颖性和多样性指标，或在冷启动时可能会表现出更好的性能，这有待进行进一步的实验研究。

在未来的研究中，还将使用分布式大数据框架处理海量数据，以验证文中模型在海量数据和分布式架构下的有效性。

参考文献：

[1] DZIUGAITE，GINTARE K，ROY D M. Neural network matrix factorization [J/OL]. [2015-12-31]. https://arxiv.org/abs/1511.06443.

[2] HE Xiangnan，LIAO Lizi，ZHANG Hanwang. Neural collaborative filtering [C] //Proceedings of the 26th International Confe-rence on World Wide Web. [S.l.]：ACM，2017：173-182.

[3] YANG C，BAI Lanxiao，ZHANG Chao. Bridging collaborative filtering and semi-supervised learning：a neural approach for poi recommendation [C] //Sigkdd International Conference on Knowledge Discovery & Data Mining. [S.l.]：ACM，2017：1245-1254.

[4] CATHERINE R，WILLIAM C. Transnets：learning to transform for recommendation [C] //Procee-dings of the Eleventh ACM Conference on Recommender Systems. [S.l.]：ACM，2017：288-296.

[5] ZHENG L，NOROOZI V，YU P S. Joint deep modeling of users and items using reviews for recommen-dation [C] //The Tenth ACM International Conference. [S.l.]：ACM，2017：425-434.

[6] WU Y，DUBOIS C，ZHENG A X. Collaborative denoising auto-encoders for top-n recommender systems [C] //The Ninth ACM International Conference. [S.l.]：ACM，2016：153-162.

[7] BANSAL T，BELANGER D，MCCALLUM A. Ask the GRU：multi-task learning for deep text recommen-dations [J/OL]. [2016-12-31]. https://arxiv.org/abs/1609.02116.

[8] WANG C，BLEI D M. Collaborative topic modeling for recommending scientific articles [C] //Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]：ACM，2011：448-456.

[9] WANG H，WANG N，YEUNG D Y. Collaborative deep learning for recommender systems [J]. OALib Journal，2014：1409.2944

[10] HE R，MCAULEY J. VBPR：visual bayesian personalized ranking from implicit feedback [C] //AAAI'16 Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. [S.l.]：ACM，2015：144-150.

[11] HU G N，DAI X Y. Integrating reviews into personalized ranking for cold start recommendation [C] //PAKDD 2017：Advances in Knowledge Discovery and Data Mining. [S.l.]：Springer International Publishing，2017：708-720.

[12] MCAULEY J，LESKOVEC J. Hidden factors and hidden topics：understanding rating dimensions with review text [C] //Proceedings of the 7th ACM conference on Recommender systems. [S.l.]：ACM，2013：165-172.

[13] CHELLIAH M，SARKAR S. Product recommendations enhanced with reviews [C] //Proceedings of the Eleventh ACM Conference on Recommender Systems. [S.l.]：ACM，2017：398-399.

[14] HU H. LCMR：local and centralized memories for collaborative filtering with unstructured text [EB/OL]. [2018-12-31]. https://arxiv.org/abs/1804.06201.

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

[15] EBESU T, SHEN B, FANG Y. Collaborative memory network for recommendation systems [C] // The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. [S. l.]: ACM, 2018: 515-524.

[16] ZHANG S, YAO L, SUN A. Deep learning based recommender system: a survey and new perspectives [J]. ACM Computing Surveys (CSUR), 2019, 52 (1): 5

[17] ZHENG L, NOROOZI V, YU P S. Joint deep modeling of users and items using reviews for recommendation [C] // Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. [S. l.]: ACM, 2017: 425-434.

[18] KIM Y. Convolutional neural networks for sentence classification [J]. OALib Journal, 2014: 1408.5882.

[19] HYUN D, PARK C, YANG M C. Review sentiment-guided scalable deep recommender system [C] // The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. [S. l.]: ACM, 2018: 965-968.

[20] ZHANG Y, YANG Q. A survey on multi-task learning [J/OL]. [2017-12-31]. https://arxiv.org/abs/1707.08114.

[21] SEO S, HUANG J, YANG H. Representation learning of users and items for review rating prediction using attention-based convolutional neural network [C] // 3rd International Workshop on Machine Learning Methods for Recommender Systems. [S. l.]: ACM, 2017: 297-305.

[22] MISRA I, SHRIVASTAVA A, GUPTA A. Cross-stitch networks for multitask learning [C] // IEEE Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE, 2016: 3994-4003.

[23] KHAN M M, IBRAHIM R, GHANI I. Cross domain recommender systems: a systematic literature review [J]. ACM Computing Surveys (CSUR), 2017, 50 (3): 36.

[24] VASWANI A, SHAZEER N, PARMAR N. Attention is all you need [J/OL]. [2017-12-31]. https://arxiv.org/abs/1706.03762.

[25] HE X, LIAO L, ZHANG H. Neural collaborative filtering [C] // Proceedings of the 26th International Conference on World Wide Web. [S. l.]: International World Wide Web Conferences Steering Committee, 2017: 173-182.

Improved Transfer Learning Algorithm Based on Cross-domain in Recommendation System

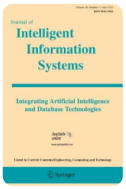
ZHANG Ziye[1] *LI Mingchang*[1] *LIANG Lingrui*[1] *ZHANG Minghua*[1] *XIE Xianfen*[2] *GU Wanrong*[3]

(1. School of Mathematics, South China University of Technology, Guangzhou 510640, Guangdong, China;
2. School of Economy, Jinan University, Guangzhou 510632, Guangdong, China;
3. School of Mathematics and Information, South China Agricultural University, Guangzhou 510642, Guangdong, China)

Abstract: In recommendation system, the single-domain recommendation algorithm faces many practical problems. The main collaborative filtering algorithms only use the information of the interaction between users and items, and can not avoid the problem of data sparsity in practical applications. This study focuses on the fusion of information from other fields to deal with sparse data, and solve the problem that it is difficult to find effective association from cross-fields. An associative feeling network model based on the deep neural network was proposed. It realizes transfer learning by mining the content information and information of items to obtain the optimal feature association, thus the accuracy of recommendation was optimized. The algorithm can deal with the problem of data sparsity in the recommendation system. Compared with some recently proposed algorithms, it shows an improved performance (15% ~20%) and better expansibility.

Key words: recommendation system; data mining; machine learning; neural network

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

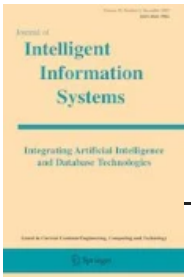


Journal of Intelligent Information Systems

Publishing model: **Hybrid**

← Back to Overview

Search all Journal of Intelligent Information Systems articles →

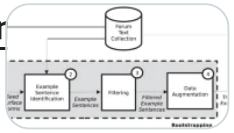


Volume 59, Issue 3

December 2022

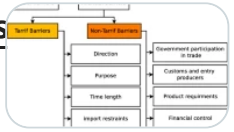
12 articles in this issue

A distantly supervised approach for recognizing product r user-generated content



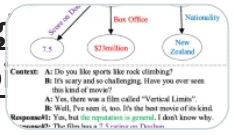
OriginalPaper | 27 May 2022 | Pages: 543 – 566

A new knowledge discovery approach for mining business barriers



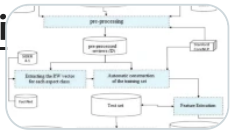
OriginalPaper | Open access | 30 May 2022 | Pages: 567 – 590

Improving knowledge-based dialogue generation throug knowledge selection and knowledge selection-guided po



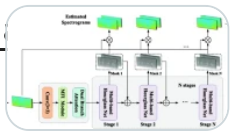
OriginalPaper | 02 June 2022 | Pages: 591 – 611

Aspect extraction and classification for sentiment analysi reviews



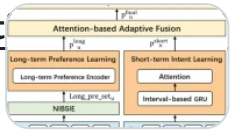
OriginalPaper | 13 June 2022 | Pages: 613 – 633

Multi-stage music separation network with dual-branch hybrid convolution



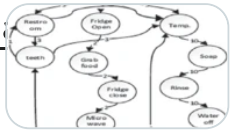
OriginalPaper | 16 June 2022 | Pages: 635 – 656

Two-stage sequential recommendation for side informat long-term and short-term preferences modeling



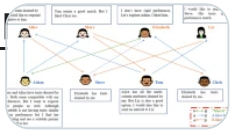
OriginalPaper | 24 June 2022 | Pages: 657 – 677

A data-driven situation-aware framework for predictive smart environments



OriginalPaper | 25 June 2022 | Pages: 679 – 704

A contextual-bandit approach for multifaceted reciproca recommendations in online dating



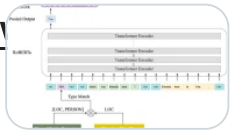
OriginalPaper | 25 June 2022 | Pages: 705 – 731

GIAD-ST: Detecting anomalies in human monitoring base generative inpainting via self-supervised multi-task lear



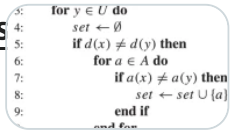
OriginalPaper | 30 June 2022 | Pages: 733 – 754

Entity-aware answer sentence selection for question ans
transformer-based language models



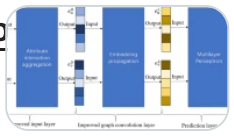
OriginalPaper | 09 July 2022 | Pages: 755 – 777

FPGA supported rough set reduct calculation for big datas



OriginalPaper | 11 July 2022 | Pages: 779 – 799

An improved recommendation based on graph convolutio



OriginalPaper | 11 July 2022 | Pages: 801 – 823

For authors

Submission guidelines →

Language editing services ↗

Ethics and disclosures →

How to publish with us →

Open Access fees and funding →

Contact the journal →

Journal of Intelligent Information Systems (2022) 59:801–823
https://doi.org/10.1007/s10844-022-00727-3

An improved recommendation based on graph convolutional network

Yichen He[1] · Yijun Mao[1] · Xianfen Xie[2] · Wanrong Gu[1]

Received: 8 June 2022 / Revised: 28 June 2022 / Accepted: 1 July 2022 /
Published online: 11 July 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract
Graph convolutional network is a recently developed artificial neural network method commonly used in recommendation system research. This paper points out three shortcomings of existing recommendation systems based on the graph convolutional network. 1. Existing models that take the one-hot encoding based on node ordinal numbers in the graph or encoding based on original entity attributes as input may not fully utilize the information carried by the attribute interactions. 2. Previous models update the node embeddings only by the first-order neighbors in the graph convolution layer, which is easily affected by noise. 3. Existing models do not take into account differences in user opinions. We propose an improved graph convolutional network-based collaborative filtering model to address these drawbacks. We identify inner and cross interaction between user attributes and item attributes, and then we take the vector representations of aggregated attributes graph as input. In the convolutional layer, we aggregate the second-order collaborative signals and incorporate the different user opinions. The experiments on three public datasets show that our model outperforms state-of-the-art models.

Keywords Graph convolutional network · High-order collaborative signal · Collaborative filtering · Recommendation system

✉ Wanrong Gu
guwanrong@scau.edu.cn

Yichen He
heyichen666@hotmail.com

Yijun Mao
yijunmao@163.com

Xianfen Xie
txiexianfen2009@jnu.edu.cn

[1] School of Mathematics and Information, South China Agricultural University, Tianhe, Guangzhou 510642, Guangdong, China

[2] School of Economy, Jinan University, Tianhe, Guangzhou 510632, Guangdong, China

1 Introduction

Collaborative Filtering (CF) is one of the most important algorithms in recommendation systems. The recommendation algorithm based on collaborative filtering has been scholars' most extensive and profound research (Kluver et al., 2018; Jalili et al., 2018). Collaborative filtering can be divided into memory-based and model-based collaborative filtering, depending on whether a machine learning model is used or not. Compared with the former, which relies on the similarity function, has high memory occupation, and is hard to discover nonlinear and implicit connections between entities, the model-based collaborative filtering uses matrix factorization (MF) (Mehta & Rana, 2017), Markov Chain (MC) (Quadrana et al., 2018), deep neural network (Liu & Wu, 2017) and other models to learn user and item embeddings from user-item interactions, and becomes the most widely studied recommendation algorithm.

In recent years, Graph Convolutional Networks (GCN) have emerged as the hottest algorithms in current recommendation systems due to their powerful ability to learn about entities and their relations on graph structures. GCN uses spectral graph theory to define graph convolution operations, which propagate and aggregate messages from neighboring nodes on the topological graph. GCN has the advantage of learning entity embedding by using both entity features and graph structure features. It extends the excellent feature learning and representation capabilities of traditional Convolutional Neural Networks (CNN) to non-regular graph data. The entities and their interactions in the recommendation scenario can be naturally represented by a topological graph, which makes it very suitable for implementing graph-based recommendation tasks based on the GCN model.

Due to the powerful feature representation learning capability, GCN-based recommendation models can better exploit the implicit and nonlinear collaborative signals on entity interaction graphs and usually obtain better accuracy than traditional collaborative recommendation models. Although graph convolutional networks have been recognized to have significant advantages in recommendation systems, the proposed GCN-based recommendation models usually neglect three aspects that need further improvement.

- A key point in designing a graph convolutional network is what information is used as the input to the network. Current graph convolutional networks take one or more of the following four as inputs: 1. The attribute information of the nodes in the graph. 2. Degrees of the nodes in the graph. 3. One-hot encodings of the node ordinal numbers in the graph. 4. Constant vectors of all 1. However, we argue that they may not be sufficient to use the important semantic features and the attribute interactions of entities, which are not conducive to accurately learning the embedding of the node.
- The graph convolution layers of most models do not take full advantage of the collaborative signals of higher-order neighbors. In contrast, Xu et al. (Xu et al., 2018) show that higher-order neighbours' collaborative signals are also important for learning node embeddings. Several models stack multiple graph convolutional layers in the GCN framework to implicitly pass higher-order collaborative signals. However, they still update node embeddings with only first-order neighbors inside their convolutional layers. These approaches are not only difficult to control the signal strength when aggregating higher-order signals but also easily affected by noise. (see the analysis in Section 3.3 for details).
- The convolutional layers of existing models pay little attention to the influence of different user opinions on the recommendation results when propagating and aggregating

collaborative signals. The ratings represent the user opinions in recommendation systems. For example, if a user gives ratings of 2 and 5 to items i_1 and i_2 respectively, he obviously expresses a more "like" opinion on i_2. However, most existing models ignore the difference in user opinions when aggregating the first-order collaborative signals. We note that the GCMC model (Berg et al., 2017) attempts to separately aggregate collaborative signals at different rating levels. However, it still does not consider the differences in user opinions when fusing signals from multiple levels.

We design an Improved Graph Convolutional Network based Collaborative Filtering (IGCN-CF) recommendation model for rating prediction to address the above shortcomings. In this paper, we use the interaction of entity attributes, the collaborative signals of second-order neighbors, and the user opinions to improve the accuracy of the graph convolutional network in learning node embedding.

We summarize the main contributions of our work as follows:

- We emphasize the importance of considering the entity's attributes on the graph convolution network, and categorize the attribute interactions into inner and cross interactions. Finally, we take the vector representations of aggregated attribute vectors as input in the IGCN-CF model.
- We propose a method to compute the second-order collaborative signals and their strength from the adjacency matrix, which uses to update the node embeddings in the graph convolution layer.
- We give a simple method to integrate user opinions into the first-order collaborative signals aggregation process, which facilitates learning of the node embeddings with user preferences.
- We propose an improved collaborative recommendation model based on the graph convolutional network, and its performance advantages are verified on several public datasets.

2 Related works

2.1 Implicit feedback-based recommendation methods

The Latent Factor Model (LFM) is the most widely used algorithm in collaborative filtering. The method represents the users' ratings of items as a matrix, then mines the low-dimensional latent feature space by factorizing the matrix, and finally re-represents the users and items on the low-dimensional space. The model finally uses the inner product between the embedding vectors of users and items to characterize the correlation between them. Most efforts to improve the embedding function focus on incorporating side information. Such as Li et al. (Li et al., 2014) introduced users' purchase, search and browsing records based on the traditional one-class collaborative filtering model. They used the added user's features to improve the accuracy of recommendations. Xin et al. (Dong et al., 2017) proposed a hybrid model, which is a variant of the stacked denoising auto-encoder. The model can efficiently incorporate users' and items' side information into latent factors. Guo et al. (Guo et al., 2019) proposed a collaborative filtering model that considers emotion and trust. The model excludes the malicious users from neighbors according to the degree of emotional consistency and obtains the recommendation list based on the trust

relationships of target users. Similarly, some scholars incorporated social relations (Wang et al., 2017) and external knowledge graphs (Wang et al., 2019; Ren et al., 2021) into the collaborative model. While LFM uses the inner product between the embedding vectors of users and items to fit the user-item interactions, its linear characteristic makes it not enough to reveal the complicated non-linear relationships between users and items (Hsieh et al., 2017).

2.2 Deep learning-based recommendation methods

Scholars improve the interaction function through deep learning (Liu & He, 2022; Fu et al., 2019) to capture the non-linear relationships between users and items. For instance, Guo et al. (Guo et al., 2017) and He et al. (He & Chua, 2017) used a Multilayer Perceptron (MLP) to model the interactions. Tay et al.(Tay et al., 2018) used Euclidean distance to represent the strength of the interaction. Although deep learning has significant advantages in recommendation systems, we argue that embedding functions of existing works are insufficient to yield optimal embeddings for collaborative filtering. Summarizing existing deep learning methods, the embedding function transforms the entities' features (e.g. ID and attributes) into vectors. Therefore, they ignore the attribute interactions, which are not conducive to learning the node embeddings accurately.

2.3 Graph convolutional networks

Graph convolutional network is one of the deep learning methods, with rich results in theory and application. With the development of GCN, several recommendation models based on graph convolutional networks have been proposed successively. Berg et al. (Berg et al., 2017) implemented a GCMC (Graph Convolutional Matrix Completion) recommendation model using GCN as an encoder in a graph self-coding framework, which uses a single graph convolutional layer to aggregate the collaborative signals of first-order neighbors on the user-item bipartite graph. Wang et al. (Wang et al., 2019) proposed an NGCF (Neural Graph Collaborative Filtering) model, which stacks multiple graph convolutional layers to achieve the implicit transmission of collaborative signals of higher-order neighbors on the bipartite graph, and finally aggregates the node embeddings in all layers for the recommendation. NGCF model compensates for the weakness of the GCMC model, which can only utilize the first-order collaborative signals. He et al. (He et al., 2020) argued that the NGCF model takes One-hot encodings with no concrete semantics besides being an identifier as input, which will increase the difficulty of model training, resulting in reduced prediction performance. They proposed a light GCN model, which removes feature transformation, neighborhood aggregation, and nonlinear activation in the embedding propagation layers. Much effort has been devoted to improving the efficiency of graph convolution operations (Chen et al., 2020), using attention mechanisms (Song et al., 2019; Feng et al., 2019), incorporating side information (Fan et al., 2019; Wang et al., 2019; Hui et al., 2022), improving model scalability (Ying et al., 2018; Wang et al., 2018), and overcoming the cold-start problem (Zhang et al., 2019) to improve the GCN-based models. We argue that existing works do not take full advantage of the collaborative signals of higher-order neighbors. For example, GCMC model only utilizes the signals of first-order neighbors to update the node embeddings. The GraphRec (Fan et al., 2019) model aggregates the collaborative signals of friends directly using social relationships, which does not exploit the higher-order signals of items and requires additional side information. Compared to other

models that implicitly pass higher-order collaborative signals by stacking multiple graph convolutions, our model uses second-order collaborative signals to update the node embedding inside the graph convolution layer.

3 Problem statement & background

3.1 Problem statement

In the recommendation systems, denote by $R \in \mathbb{R}^{m \times n}$ the sparse rating matrix, which contains few non-zero elements $r_{u,i} \in \{1, 2, ..., T\}$. $r_{u,i}$ represent user u rating of item i, with T levels. Denote by $X_u \in \mathbb{R}^{m \times d_u}$ the entity characteristic of users, and by $X_v \in \mathbb{R}^{m \times d_v}$ the entity characteristic of items, where m and n is the number of user set $\mathcal{U}$ and item set $\mathcal{V}$; d_u and d_v is the dimension of the characteristic vector of users and items. The goal of the recommendation problem is to design a model to predict ratings of items that have not been rated and then realize the recommendation.

In graph neural network-based models, users, items and their rating relationships are usually considered as undirected bipartite graphs $\mathcal{G} = \langle E, A \rangle$, where $E = \mathcal{U} \cup \mathcal{V}$ is the node set in the graph, and adjacent matrix $A \in \mathbb{R}^{(m+n) \times (m+n)}$ represents whether there is an interaction between nodes, which can be obtained directly from the rating matrix R. At this point, the rating prediction problem can be considered as a prediction of the unknown links on the bipartite graph $\mathcal{G}$.

3.2 GCN-based recommendation model

Recommendation models based on graph convolutional networks use the idea of message passing to propagate and aggregate collaborative signals over the bipartite graph, and learn the node embeddings for recommendation prediction. The model structure usually consists of the input layer, graph convolution layer, and prediction layer.

3.2.1 Input layer

The entity nodes on the graph $\mathcal{G}$ are encoded to obtain their initial low-dimensional embedding vector e^0, such as:

$$e^0 = Enc(x) \tag{1}$$

where x is the information of entity nodes, which can be represented by entity original characteristics or one-hot encoding vectors based on node ordinal numbers; e^0 is the initial embedding vector after encoding, which can be subdivided into user embedding e_u^0 and item embedding e_v^0. $Enc(\cdot)$ is the encoding function.

3.2.2 Convolution layer

The convolution layer is also known as the "embedding propagation layer". It uses the structural features of the bipartite graph to propagate and aggregate the collaborative signals of neighboring nodes on the graph to achieve the node embedding update. Its main operations include collaborative signals construction and node embeddings update. We

take a user u rating of an item v as an example, the rating behavior reflects a certain degree of user preference for the item. Therefore, the first-order collaborative signal propagated by this behavior can be constructed as:

$$s_{u \leftarrow v} = f(e_u, e_v, p_{u,v}, \Theta^f) \tag{2}$$

where $p_{u,v} = 1/\sqrt{|\mathcal{N}_u| \cdot |\mathcal{N}_v|}$ is the signal strength coefficient. Denote by $\mathcal{N}_u$ and $\mathcal{N}_v$ the set of directly adjacent first-order neighbors of user u and item v, where $|\mathcal{N}_u|$ and $|\mathcal{N}_v|$ is the node degree. $f(\cdot)$ is the encoding function of collaborative signals, where Θ^f are all parameters to be learned in the function.

The embedding update operation aggregates all collaborative signals of first-order neighbors on the bipartite graph. The update of the user embeddings can be expressed as:

$$e_u^{update} = \mathrm{g}(e_u, \{s_{u \leftarrow v} | v \in \mathcal{N}_u\}) \tag{3}$$

where aggregation function $\mathrm{g}(\cdot)$ can be implemented by weighted average, max pooling, LSTM network, etc.

Similarly, when item v receives a rating from user u, it indicates that the item's quality matches the user's preference to some extent. Therefore, the embedding of the item can be updated using the user collaborative signals as:

$$s_{v \leftarrow u} = f(e_v, e_u, p_{u,v}, \Theta^f) \tag{4}$$

$$e_v^{update} = \mathrm{g}(e_v, \{s_{v \leftarrow u} | u \in \mathcal{N}_v\}) \tag{5}$$

In this way, the graph convolution layer uses the first-order collaborative signals to achieve a single update of the entity node embeddings. In addition, many GCN-based recommendation models stack L graph convolution layers to iteratively refine the node embedding, which get $e_u^1, e_u^2, \ldots, e_u^L$ and $e_v^1, e_v^2, \ldots, e_v^L$ in order. Our method also implicitly realizes the layer-by-layer propagation of higher-order collaborative signals.

3.2.3 Prediction layer

The prediction layer uses the final node embedding e_u^L and e_v^L to predict the unknown ratings between any user u and item v, as:

$$\hat{r}_{u,v} = h(e_u^L, e_v^L, \Theta^h) \tag{6}$$

where prediction function $h(\cdot)$ is usually implemented by a bilinear decoding function or a multilayer perceptron network; Θ^h are all parameters in the function.

3.3 Problem analysis

As described in Section 1, a key point in designing a graph convolutional network is what information is used as the input to the network. The existing encoding functions do not utilize entity attributes or analyze all attribute interactions equally, which encode all user and item attributes as vectors without distinguishing their interactions. For example, suppose there is a user with attributes <Female,20-24> and an item with attributes <Romance,Comedy>. Existing encoding functions ignore the interaction, such as "whether

female users like romance movies" (<Female,Romance>). We argue that they do not take full advantage of the entity's attribute information.

On the one hand, as described in Section 3.2, the collaborative signals aggregation (3) and (5) only use collaborative signals from directly interacting first-order neighbors and themselves, so they ignore the higher-order signals inside the graph convolution layer. However, according to the basic idea of collaborative filtering, user behaviors or preferences are also influenced by other similar users who can be considered as higher-order neighbors on the bipartite graph. Therefore, it is beneficial to aggregate higher-order signals in the graph convolution layer. Recently, Xu et al. (Xu et al., 2018) have pointed out the value of higher-order signals for optimizing node embeddings.

Several GCN-based recommendation models stack multiple graph convolutional layers to pass higher-order collaborative signals implicitly. However, they are not only difficult to control the strength when aggregating higher-order signals but also easily affected by noise. For example, suppose there is a connected path $u_0 \rightarrow v_1 \rightarrow u_2 \rightarrow v_3 \rightarrow u_4$ on the bipartite graph, where user u_4 rate item v_3 for accidental reasons. Although it is not similar to the preference of the target user u_0, this rating behaviour will still result in the collaborative signal of u_4 being implicitly passed backwards along with the stacked four convolutional layers to u_0, rather than be considered as a noisy signal being suppressed or eliminated.

On the other hand, in the aggregation (3) and (5), all the first-order neighbor signals are treated equally without considering the difference in the user's opinions (i.e. rating value). However, recent researches (Fan et al., 2019; Xiang et al., 2010) point out that high-rated items are more reflective of user preferences than low-rated items and should play a more significant role in signal aggregation operations.

4 Our approach

In this section, we describe an end-to-end IGCN-CF model to address the shortcomings of the GCN-based recommendation model as described in Section 3.3. First, we improve the coding function of the input layer, fuse the inner interaction and cross interaction of the entities, and finally take the aggregated vectors with the characteristic information as input. Then, we introduce second-order collaborative signals and user opinions inside the graph convolution layer, which learn the node embeddings more effectively. Finally, we use MLP to achieve predictions of unknown ratings. The basic structure is shown in Fig. 1, including the improved input layer, the improved graph convolution layer and the prediction layer.

4.1 Input layer

Inspired by (Su et al., 2021), we fuse the inner interaction and cross interaction of the entities and take the aggregated vectors as input. Figure. 2 shows an overview of the input layer. First, we construct the user attribute graphs and item attribute graphs to represent the users and items. The attributes are treated as nodes in the attribute graphs, and each pair of nodes are connected as edges. Therefore, the attribute graphs are complete graphs. For example, if a user is male and 20-24 years old, in the attribute graph of this user, we take "male" and "20-24" as nodes, and two nodes are connected with an edge. Donate by u_i^U the node i embedding in a user attribute graph and by u_i^I the node i embedding in an item attribute graph. Due to the symmetric modelling of user attribute graphs and item attribute graphs, we omit the superscript U and I in the following section.

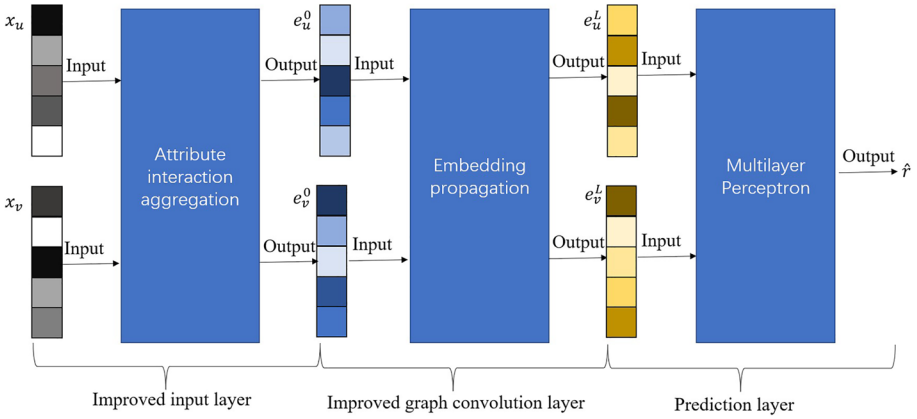


Fig. 1 The basic structure of IGCN-CF

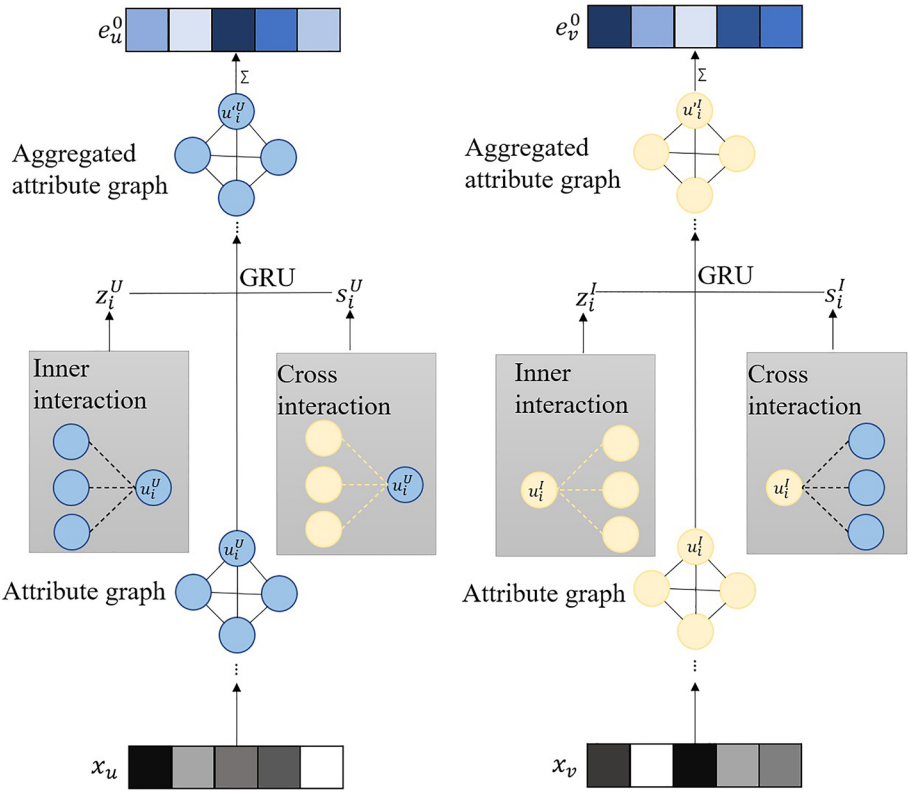


Fig. 2 An overview of the input layer

4.1.1 Inner interaction

Inner interactions are used for user (item) characteristics learning. We use the $MLP \in \mathbb{R}^{2\times d} \rightarrow \mathbb{R}^d$ to model the inner interactions:

$$z_{i,j} = MLP(u_i, u_j) \tag{7}$$

where z_{ij} are the inner interaction results of node pair i and j, u_i and u_j is the embedding of node i and j.The same modelling is performed for each node pair in the attribute graph, and the final results of the modelling are aggregated with the element-wise sum:

$$z_i = \sum_{j \in N_i} z_{ij} \tag{8}$$

where z_i is the aggregated inner interaction results of node i, and N_i is a set of neighbors of node i.

Moreover, inner interactions are employed to capture user (item) characteristics and are intrinsically complicated. Note that in contrast to cross interaction model results that reveal similarity (will be discussed in Section 4.1.2), a high inner interaction result does not imply that the two attributes are similar when learning the characteristics of a user (an item). As a result, we use a neural approach that non-linearly to model the two attributes.

4.1.2 Cross interaction

Cross interactions describe the dependency between the user attributes and the item attributes. We expect a high interaction result between an attribute u_i^U and an attribute u_j^I if u_i^U has a strong preference for u_j^I. For example, if female users prefer romance movies, the interaction result between the attributes "Female" and "Romance" should be high. Moreover, if a user attribute has a strong preference for an item attribute, their embeddings should be similar after training in collaborative filtering. Inspired by (He & Chua, 2017), we use Bi-interaction to model the cross interaction:

$$s_{ij} = u_i \odot \hat{u}_j \tag{9}$$

where s_{ij} are the cross interaction results of node pair i and j, u_i is the embedding of node i in one graph, and $\hat{u}_j$ is the embedding of node j in the other graph. The element-wise product $\odot$ makes sure s_{ij} is high if node i has a high preference of node j. Similar to inner interaction, we aggregate the results with the element-wise sum:

$$s_i = \sum_{j \in \hat{V}} s_{ij} \tag{10}$$

where s_i is the aggregated cross interaction results of node i, and $\hat{V}$ is a set of nodes in the other attribute graph.

4.1.3 Information fusing

We can get the final embeddings by fusing the initial node embeddings u_i, the inner interaction results z_i and the cross interaction results s_i. The fusing function reference document (Su et al., 2021), we use Gated Recurrent Unit ($GRU \in \mathbb{R}^{3\times d} \rightarrow \mathbb{R}^d$) to get the fused node representation:

$$u_i' = GRU(u_i, z_i, s_i) \tag{11}$$

Then we aggregate the node representations with the element-wise sum:

$$e^0 = \sum_{i \in V} u_i' \tag{12}$$

where e^0 are the initial input of the graph convolution network in this paper.

Note that the user or item attributes may not be available in some cases. In this case, our model can also be used. When the user and item attributes are unavailable, the attribute graph becomes a single node (user/item ID). Then there is no inner interaction modelling, simply cross interaction modelling between the user ID and the item ID, i.e. $ui' = GRU(u_i,s_i)$, where u_i is the node representation of entity ID.

4.2 Graph convolution layer

Similar to the NGCF model, the model in this paper achieves layer-by-layer refinement of the node embedding representation by stacking L graph convolution layers. We improve the graph convolution layer by aggregating the second-order collaborative signals and user opinions. This section takes the $l \in \{1, 2, \ldots, L\}$-th graph convolution layer as an example to introduce the main idea of our models. Figure. 3 shows an overview of the graph convolution layer.

4.2.1 Second-order collaborative signals construction

Let $u_i \rightarrow v_k \rightarrow u_j \rightarrow \ldots$ be any connectivity path from user u_i on the bipartite graph $\mathcal{G}$, where $i,k,j,\ldots$ represent the node number. We call v_k a first-order neighbor of u_i, u_j a second-order neighbor of u_i, and so on, according to the number of hop on the path from user u_i.

As discussed in Section 3.3, aggregating higher-order collaborative signals in the graph convolution layer is beneficial for learning node embedding representations. However, our model only uses the collaborative signals of the second-order neighbors

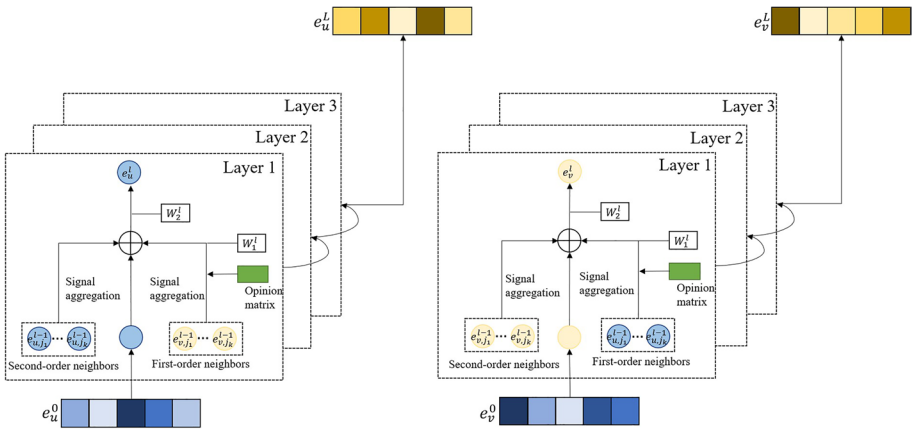


Fig. 3 An overview of the graph convolution layer

among all higher-order neighbors for two following reasons: 1. According to Xu et al. (Xu et al., 2018), the strength of the collaborative signals of higher-order neighbors on the connected path decreases rapidly with increasing order, and the noise signals tend to accumulate on the path. This assertion also explains why many GCN-based models stack only three or four convolutional layers. 2. Second-order neighbors are homogeneous entities with common interaction items (users) with the target users (items). As explained by Triadic Closure theory (Easley & Kleinberg, 2010), second-order neighbors are more similar in their preferences to the target entity than higher-order neighbors, and the more common interaction entities that second-order neighbors have with the target entity, the more similar they are. Moreover, the experimental results in Section 5 verify that adding fourth-order collaborative signals cannot significantly improve the performance of the recommended model.

Inspired by the document (Wang et al., 2006) in terms of recommended entity similarity: the more similar neighbors transmit more signals, and the neighbors with more common interaction entities transmit more signals. We propose a method to compute the second-order collaborative signals and their strength. Let u_j be the second-order neighbor of target user u_i on the graph $\mathcal{G}$, $u_i \rightarrow v_k \rightarrow u_j$ be any connected path of length two between them. We define the second-order signal propagating from u_j to u_i as:

$$s^l_{i\leftarrow j} = \sum_{k\in\mathcal{N}_i\cap\mathcal{N}_j} p_{i,k,j}(e^{l-1}_{u,j} + e^{l-1}_{u,i} \odot e^{l-1}_{u,j}) \tag{13}$$

where $e^{l-1}_{u,j}$ denotes the embedding of user u_j in the $l-1$-th layer, $\odot$ denotes the element-wise product, $\mathcal{N}_i \cap \mathcal{N}_j$ denotes the set of commom interaction items between u_i and u_j, and $p_{i,k,j}$ denotes the signal strength coefficient of the second-order collaborative signals in the network, which is related to the degree of nodes on the path. In this paper, we set $p_{i,k,j} = 1/(\sqrt{|\mathcal{N}_i|} \cdot |\mathcal{N}_k| \cdot \sqrt{|\mathcal{N}_j|})$. Note that $e^{l-1}_{u,i} \odot e^{l-1}_{u,j}$ also plays a role in regulating the strength of collaborative signals of neighboring nodes, so that similar neighbors propagate more second-order collaborative signals. For example, if a neighbor u_j and a target user u_i have the similar embedding, since the vectors have the same direction, the product term will enhance their collaborative signal, otherwise, the effect will be weak.

The analysis shows that (13) has a good suppression effect on the noise signal. Suppose user u_j does not have similar preferences to target user u_i, but becomes a second-order neighbor of u_i only through a chance rating operation. However, since their node embeddings are not similar, resulting in weak signal enhancement of the element-wise product term, and considering that their common interaction items $\mathcal{N}_i \cap \mathcal{N}_j$ are very few, the noise collaborative signals propagated from u_j to u_i are also very limited.

Denote by $\mathcal{N}^{(2)}_i$ all the second-order neighbors of user u_i, the aggregated second-order collaborative signal of u_i is:

$$s^{l(2)}_{u,i} = \sum_{j\in\mathcal{N}^{(2)}_i} s^l_{i\leftarrow j} = \sum_{j\in\mathcal{U},j\neq i} \sum_{k\in\mathcal{N}_i\cap\mathcal{N}_j} \frac{e^{l-1}_{u,j} + e^{l-1}_{u,i} \odot e^{l-1}_{u,j}}{\sqrt{|\mathcal{N}_i|} \cdot |\mathcal{N}_k| \cdot \sqrt{|\mathcal{N}_j|}} \tag{14}$$

Symmetrically, it is easy to get the aggregated second-order collaborative signal $s^{l(2)}_{v,i}$ of item v_i in the l-th layer. Here we will not go into details.

Denote by $E^{l-1} \in \mathbb{R}^{(m+n)\times d}$ the matrix of all embedding vectors in the $(l-1)$-th layer, and by $S^{l(2)} \in \mathbb{R}^{(m+n)\times d}$ the signal matrix aggregated by second-order collaborative signals of all nodes in the l-th layer, where row vectors consist of $s^{l(2)}_{u,i}$ or $s^{l(2)}_{v,i}$. Based on the

spectral graph theory(Bruna et al., 2013), the concise form of the second-order collaborative signal matrix $S^{l(2)}$ is:

$$S^{l(2)} = \bar{\mathcal{L}}^{(2)} E^{l-1} + (\bar{\mathcal{L}}^{(2)} E^{l-1}) \odot E^{l-1} \tag{15}$$

where $\bar{\mathscr{L}}^{(2)}$ is the de-diagonalization second-order Laplacian matrix, $\bar{\mathscr{L}}^{(2)} = \mathscr{L}^{(2)} - diag(\mathscr{L}^{(2)})$. $\mathscr{L}^{(2)}$ is the second-order Laplacian matrix which represents the second-order signal strength based on the node degree and is the matrix form of the signal strength coefficient $p_{i,k,j}$:

$$\mathcal{L}^{(2)} = (D^{-\frac{1}{2}} \cdot A \cdot D^{-\frac{1}{2}})^2 \tag{16}$$

where D is the diagonal matrix consists of node degrees; A is the adjacency matrix of graph $\mathcal{G}$. In this paper, $A = \begin{pmatrix} 0 & z(R) \\ z(R^T) & 0 \end{pmatrix}$, where $z(\cdot)$ is a function that resets the non-zero elements of the matrix to one, R is the rating matrix.

4.2.2 First-order collaborative signals with user opinions construction

Users' ratings of items indicate the existence of an interactive relationship between them, and specifically express the users' opinions. For example, in the movie recommendation, a rating of 5, 3, and 1 expresses three different opinions of users: "like it very much", "like it a little", and "do not like it". According to the document (Xiang et al., 2010), high ratings with positive views are more reflective of users' true preferences. Therefore, incorporating user opinions when aggregating the first-order collaborative signals can more accurately learn user embeddings.

We draw on the message passing mechanism of GCN and propose a first-order collaborative signal with the user opinions construction method. It denotes the aggregated first-order collaborative signal converged of user u_i as:

$$s_{u,i}^{l(1)} = \sum_{j \in \mathcal{N}_i} c_{i,j} p_{i,j} W_1^l e_{v,j}^{l-1} \tag{17}$$

where opinion coefficient $c_{i,j}$ denotes user u_i's opinion of item v_j, the values in this paper are taken as normalized ratings. Specifically, $c_{i,j} = ru_i,v_j/max(R)$; $p_{i,j}$ is the first-order collaborative signal strength, which is the same as $p_{u,v}$ in (2); $W_1^l \in \mathbb{R}^{h \times h}$ is the weight matrix to be learned in the l-th layer.

Obviously, the opinion coefficients of the higher rating items take larger values, and they transmit more collaborative signals. Therefore, this approach with opinions aggregates the first-order collaborative signals more reasonably.

Symmetrically, it is easy to get the aggregated first-order collaborative signal $s_{v,i}^{l(1)}$ of item v_i in the l-th layer. We will not go into details.

Similarly, We denote the concise form of the first-order collaborative signal matrix based on the spectral graph theory as:

$$S^{l(1)} = (C \odot \mathcal{L}^{(1)}) E^{l-1} W_1^l \tag{18}$$

where C is the normalized opinion coefficient matrix consist of $c_{i,j}$, specofocally, $C = \begin{pmatrix} 0 & \frac{R}{max(R)} \\ \frac{R^T}{max(R)} & 0 \end{pmatrix}$; $\mathscr{L}^{(1)}$ is the first-order Laplacian matrix, and $\mathscr{L}^{(1)} = (D^{-\frac{1}{2}} \cdot A \cdot D^{-\frac{1}{2}})$.

4.2.3 Node embeddings update

We aggregate the different order collaborative signals and node embeddings in the previous layer, and update the node embedding vectors in the l-th layer as:

$$E^l = ReLU((E^{l-1} \oplus S^{l(1)} \oplus S^{l(2)})W_2^l) \tag{19}$$

where $ReLU(\cdot)$ is the activation function; $\oplus$ denotes the concatenate operation; $W_2^l \in \mathbb{R}^{3h \times h}$ is the weight matrix to be learned. Note that (19) fuses the embedding of the previous layer E^{l-1} to the collaborative signals as well, which is equivalent to the self-connected signals on the bipartite graph (Chen et al., 2020), and helps to avoid the "Over-Smoothing" problem in the embedding learning process.

The IGCN-CF model can stack L graph convolution layers to achieve layer-by-layer refinement of the embedding representation according to (19). Then we can get $E^1, E^2, \ldots, E^L$.

4.3 Nonlinear prediction layer

Figure 4 shows an overview of the input layer. The final learned node embeddings e_u^L and e_v^L of the graph convolution layer are connected and input to a 2-layers MLP network to predict the unknown rating $\hat{r}_{u,v}$ between user u and item v:

$$\hat{r}_{u,v} = W_4(ReLU(W_3(e_u^L \oplus e_v^L))) \tag{20}$$

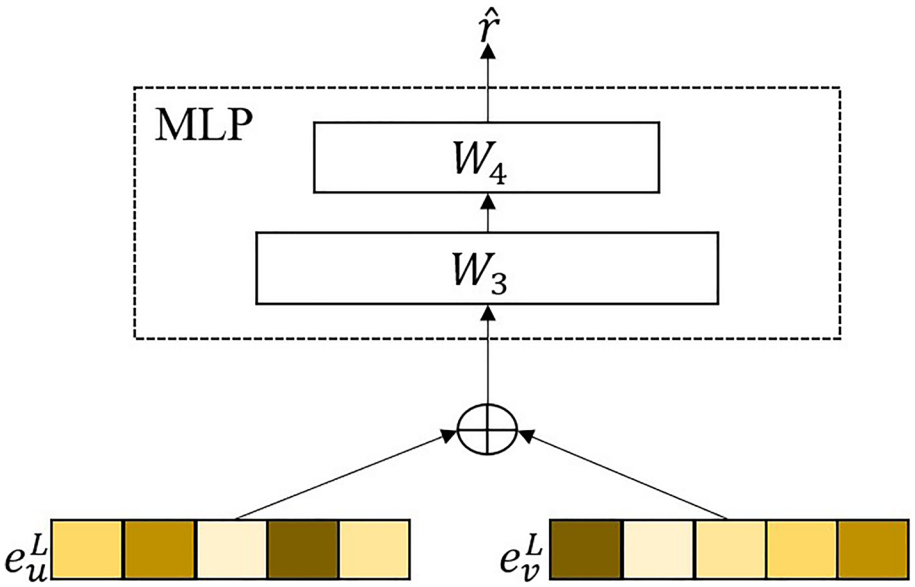


Fig. 4 An overview of the prediction layer

where $W_3 \in \mathbb{R}^{2h \times d'}$ and $W_4 \in \mathbb{R}^{d' \times 1}$ is the weight matrix of MLP to be learned; d' is the number of neurons in the first layer (hidden layer) of the MLP.

Noted that compared to some linear prediction models used in the documents (Berg et al., 2017; Wang et al., 2019; Chen et al., 2020), the nonlinear MLP model used in this paper has a stronger ability to capture the complex relationships between node embedding vectors and is more suitable for the rating prediction.

4.4 Model training

The IGCN-CF model is an end-to-end recommendation model. We give an optimized objective function based on the mean square error to train the parameters:

$$obj = \frac{1}{2|O|} \sum_{(u,v) \in O} (r_{u,v} - \hat{r}_{u,v})^2 + \lambda \|\Theta\|_2^2 \tag{21}$$

where (u,v) is the edges of ratings $r_{u,v}$ on the bipartite graph $\mathcal{G}$, and O is the set of them; $\Theta = \left\{ \left\{ W_1^l, W_2^l \right\}_{l=1}^{L}, W_3, W_4 \right\}$ is all parameters to be trained in the model; $\|\Theta\|_2^2$ represrnts the L_2 norm to regularize all parameters, where λ is the regularization coefficient.

In addition, the node dropout method used in the document (Wang et al., 2019) is also borrowed in the model training to prevent overfitting, and the dropout rate is η. Finally, we use the Adam gradient descent optimizer (Kingma & Ba, 2014) to optimize (21), and use the minibatch method to improve the model training efficiency.

4.5 Discussions

We aggregate the collaborative signals of homogeneous second-order neighbors explicitly inside the graph convolution layer of the IGCN-CF model. Theoretically, it is possible to aggregate higher-order collaborative signals. For example, when aggregating both second and fourth order collaborative signals in the bipartite graph, it is sufficient to simply replace(16) as:

$$\mathcal{L}^{(2,4)} = (D^{-\frac{1}{2}} \cdot A \cdot D^{-\frac{1}{2}})^2 + (D^{-\frac{1}{2}} \cdot A \cdot D^{-\frac{1}{2}})^4 \tag{22}$$

In addition, in the recommendation model proposed by Fan et al. (Fan et al., 2019), the collaborative signals of friends in social networks are aggregated into the graph convolution layer. Social friends are equivalent to homogeneous higher-order neighbors with high similarity in the bipartite graph, and aggregating their collaborative signals is obviously beneficial for learning the embedding representation of user nodes. By contrast, the model in this paper reasonably computes the second-order neighbors and their signal strengths by matrix operations without additional side information. Moreover, when there is side information about homogeneous node relationships in the dataset, such as social relationship matrix $Q_u \in \mathbb{R}^{m \times m}$ and item similarity matrix $Q_v \in \mathbb{R}^{n \times n}$, it is easy to incorporate them into the model by modifying (16) as:

$$\mathcal{L}^{(2,4)} = (D^{-\frac{1}{2}} \cdot A \cdot D^{-\frac{1}{2}})^2 + \alpha \begin{pmatrix} Q_u & 0 \\ 0 & Q_v \end{pmatrix} \tag{23}$$

where α is the incorporation parameter.

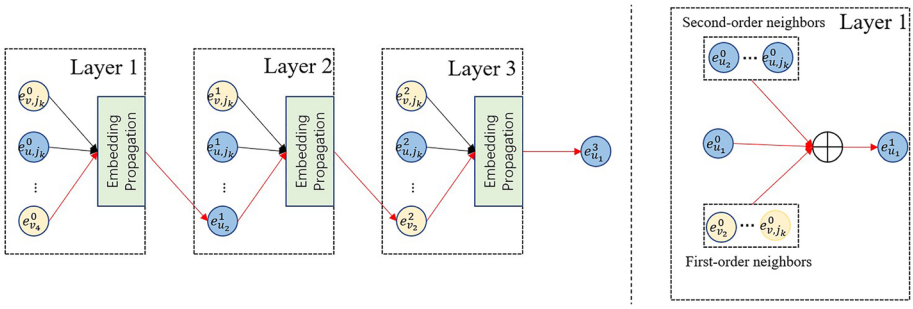


Fig. 5 Illustration of the differences between NGCF, LightGCN (left) and our model (right) in the collaborative signals utilization

Table 1 Statistics of three datasets

Dataset	Users	Items	Ratings	Density	Ratings range
MovieLens-1M	5,950	3,514	1,149,238	0.0550	1∼ 5
Book-crossing	4,873	53,168	1,050,834	0.0041	1∼ 10
Yelp2018	31,668	38,048	1,561,406	0.0013	1∼ 5

Finally, the 3-layer structure of IGCN-CF has similarities with the NGCF and LightGCN model, but they have significant differences in the input layer, collaborative signals utilization, user opinions combination, and prediction layer design. Taking Fig. 5 as an example, the red arrows show the propagation of the signals. NGCF and LightGCN stack 3 convolution layers to propagate and aggregate the higher-order collaborative signals like $u_1 \leftarrow v_2 \leftarrow u_2 \leftarrow v_4$. As discussed in Section 3.3, we argue that this approach is easily affected by noise. Thus our model propagates and aggregates the higher-order collaborative signals inside the graph convolution layer rather than stacking multiple layers. Compared with NGCF and LightGCN, our model can better control the strength of the higher-order collaborative signals and reduce the effect of the noisy signals.

5 Experiments

We conduct experiments on three real-world datasets to evaluate the proposed IGCN-CF recommendation model and compare its performance with state-of-the-art models.

5.1 Datasets

To evaluate the effectiveness of IGCN-CF, we conduct experiments on three datasets: MovieLens 1M, Book-crossing, and Yelp2018. Table 1 shows their statistic information. Below are the descriptions of the datasets:

MovieLens-1M is a dataset collected by the GroupLens Research on movie ratings, which is the most famous dataset in the recommendation system. We use the pre-processed

dataset provided in document (Su et al., 2021), which enriches the dataset with movies' other attributes, such as directors and casts from IMDB. In the dataset, the ratings greater than 3 are regarded as positive ratings, and we retain users with more than 10 positive ratings.

Book-crossing is a dataset with users' implicit and explicit ratings of books. We also use the pre-processed dataset provided in document (Su et al., 2021). In the dataset, all rated explicit ratings are regarded as positive ratings due to its sparsity, and we retain users with more than 20 positive ratings.

Yelp2018 is a dataset adopted from the 2018 edition of the Yelp challenge. We user the pre-processed dataset provided in document (He et al., 2020). In the dataset, all ratings are regarded as positive ratings, and we retain users with more than 10 positive ratings. Note that this dataset does not contain any user or item attribute information.

5.2 Experimental set-up

In the experiments, we selected four state-of-the-art recommendation models as baselines: NFM (He & Chua, 2017)[1], GCMC (Berg et al., 2017)[2], GraphRec (Fan et al., 2019)[3], NGCF (Wang et al., 2019)[4], and LightGCN (He et al., 2020)[5]. We run the codes released by the authors. NFM uses a deep neural network to replace the dot product operation of embedding vectors. In our experiments, we use Equation (21) to replace its original optimization objective so that it can be applied to rating prediction. The other four models are the prevailing GCN-based collaborative filtering recommendation models. For the NGCF and LightGCN model, we also use Equation (21) instead of their original paired BPR (Bayesian personalized ranking) loss objective function for rating prediction.

Unless otherwise specified, we use the following hyper-parameter settings: The embedding size h is fixed to 64; the number of graph convolutional layers L is fixed to 3; the regularization coefficient λ is searched in {0.01,0.05,0.1,0.5,1,5,10} by cross validation; the hidden layer parameter $d^{'}$ of MLP is 16; we use Adam as the optimization algorithm, where the learning rate is 10^{-3} and the batch size is fixed at 1024. The maximum number of epochs is 1000; the size of each batch of training data is set to 1024; the node dropout ratio η is 0.3. For all the four baselines, we use the author-provided implementations.

We randomly split each dataset into the training set, validation set and test set in the ratio of 8:1:1. We use the NDCG@K and recall@K to evaluate the effectiveness of top-K recommendation and preference ranking of our model and baseline models. By default, we set $K = 10$. All results are the average of three replicate experiments.

[1] https://github.com/hexiangnan/neural_factorization_machine
[2] https://github.com/riannevdberg/gc-mc
[3] https://github.com/wenqifan03/GraphRec-WWW19
[4] https://github.com/xiangwang1223/neural_graph_collaborative_ltering
[5] https://github.com/kuandeng/LightGCN

Table 2 Summary of the performance in comparison with baselines

	Moveilens-1M		Book-crossing		Yelp2018	
	recall	ndcg	recall	ndcg	recall	ndcg
NFM	0.1731	0.4953	0.1025	0.1768	0.1208	0.1973
GCMC	0.2108	0.5743	0.1198	0.1908	0.1393	0.2079
GraphRec	0.2251	0.6408	0.1275	0.2079	0.1533	0.2284
NGCF	0.2288	0.6375	0.1312	0.2121	0.1547	0.2329
LightGCN	0.2304	0.6477	0.1339	0.2159	0.1642	0.2445
IGCN-CF	**0.2336**	**0.6557**	**0.1359**	**0.2204**	**0.1612**	**0.2424**
%Improv.	1.39%	1.24%	1.49%	2.08%	-1.86%	-0.87%

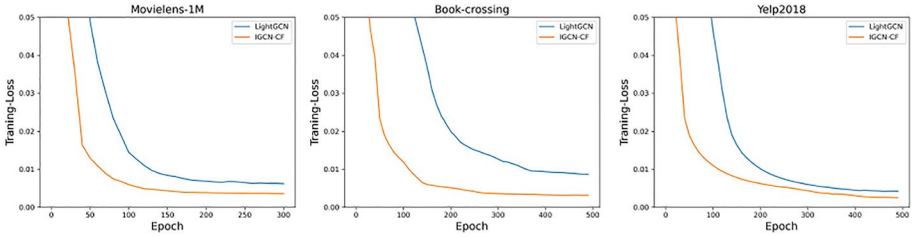


Fig. 6 Training Loss curves of IGCN-CF and LightGCN on each dataset

5.3 Performance comparison

We compare the performance of the IGCN-CF model with the comparison methods on each dataset, and their prediction performances on the test set are shown in Table 2. The numbers in bold represent the best results for each dataset, and the numbers in underline represent the best baseline results. We further compare the training loss between our model and the optimal baseline LightGCN on each dataset. The results are shown in Fig. 6.

Observing the experimental results in Table 2, four conclusions can be drawn:

- The prediction performances of the four recommendation models based on graph convolutional networks are significantly better than those of the NFM model based on matrix factorization. This result indicates that the GCN-based models can learn the embedding representations of nodes more accurately and get better prediction results due to the fusion of entity attributes of nodes and graph structure features on the topological graph.
- The GCMC model has the worst performance among the four GCN-based recommendation models, especially on Yelp2018, which is a very sparse dataset, because it uses only one graph convolutional layer and only aggregates the collaborative signals of first-order neighbors. Therefore, it has insufficient learning ability for node embedding.
- GraphRec, NGCF and LightGCN leverage higher-order collaborative signals, achieving better prediction results than the GCMC model. The difference is that the GraphRec model directly aggregates the collaborative signals of social friends, who are equivalent to similar higher-order neighbors in a bipartite graph, while the NGCF and LightGCN model implicitly propagates and aggregates higher-order collaborative signals by stack-

Table 3 Summary of the performance in comparison with baselines

	Moveilens-1M		Book-crossing		Yelp2018	
	recall	ndcg	recall	ndcg	recall	ndcg
IGCN-CF-oh	0.2292	0.6401	0.1325	0.2164	0.1556	0.2332
IGCN-CF-s4	**0.2340**	0.6553	0.1349	**0.2211**	0.1603	0.2399
IGCN-CF-ro	0.2307	0.6503	0.1353	0.2185	0.1570	0.2361
IGCN-CF-s1	0.2256	0.6130	0.1271	0.2022	0.1475	0.2277
IGCN-CF-l	0.2331	0.6527	0.1347	0.2188	0.1610	0.2417
IGCN-CF	0.2336	**0.6557**	**0.1359**	0.2204	**0.1612**	**0.2424**

ing multiple graph convolution layers. This illustrates that aggregating higher-order collaborative signals helps to improve the performance of the recommendation model.

- The IGCN-CF model in this paper achieves the best prediction performance on most of the datasets, only slightly worse than the LightGCN model on the Yelp2018. Because the Yelp2018 contains only the users' and item's IDs without any attribute information, we can not fully utilize the attribute interactions. Thus the prediction performance of our model is not optimal in this dataset. However, as shown in Fig. 6, compared with the LightGCN model, our model consistently obtains lower training loss, and the training loss of our model tends to be smooth in lower epochs, which means that our model fits the training data better than LightGCN. Our model outperforms the LightGCN model on Movielens-1M and Book-crossing but is slightly worse on Yelp2018 also indicates the significance of the attribute interaction. Our model performs better when there is sufficient attribute information about the users and the items. Moreover, we use "A/B testing" for the significance test and find that the experimental results of the IGCN-CF model are significantly better than other models with 95% confidence in most cases. The superior performance of our model can be attributed to the fact that the model leverages the attributes of entities and integrates them as inputs. In contrast, the other three GCN-based model only use one-hot encoding vectors based on node ordinal numbers as input. Furthermore, our model further exploits higher-order collaborative signals inside the graph convolution layer and introduces user opinions into the signal aggregation. In contrast, the GraphRec model only obtains the higher-order collaborative signals of users directly from social relationships without considering the higher-order signals of items; the NGCF and LightGCN model does not explicitly exploit the higher-order collaborative signals inside the graph convolutional layer and does not consider the influence of users' opinions.

5.4 Study of IGCN-CF

To further analyze the reasons for the excellent performance achieved by the IGCN-CF model, we give five variants of its model and compare their performance on the datasets, including: IGCN-CF-oh, IGCN-CF-s4, IGCN-CF-ro, IGCN-CF-s1, IGCN-CF-l. Compared with the basic model, the IGCN-CF-oh model takes one-hot encoding vectors based on node ordinal numbers as input; the IGCN-CF-s4 model employs the fourth-order collaborative signals; the IGCN-ro model removes the user opinions in (18); the IGCN-CF-s1 model removes the second-order collaborative signals in (19); the IGCN-CF-1 directly uses the inner product $(e_u^L)^T \cdot e_v^L$ instead of the nonlinear neural network in (20). The

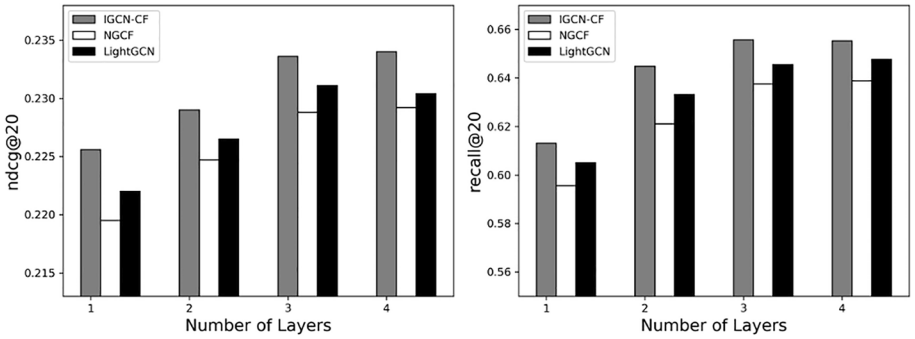


Fig. 7 Results of three models at different layer numbers on Movielens-1M

experimental parameter settings are consistent with those described in Section 5.2. The average prediction performances of three replicate experiments are summarized in Table 3.

According to Table 3, we find that:

- The prediction performance of the IGCN-CF-oh model is inferior to the basic model in all data sets. This result means that distinguishing the entity attribute interactions into inner interactions and cross interactions, taking the aggregated vectors as input helps to improve the prediction performance.
- Although the IGCN-CF-s4 model adds fourth-order collaborative signals, its prediction performance is no better than that of the basic model. This result indicates that the higher-order collaborative signals are weaker and may have noise, which is not helpful in significantly improving the performance of the recommended model. This is also the reason why the model in this paper only leverages the homogeneous second-order collaborative signals in the graph convolution layer.
- The prediction performance of the IGCN-CF-ro model without the user's opinions is lower in all datasets than in the basic model. This result shows that incorporating the user's opinions is beneficial to improving the performance of the recommendation model.
- The IGCN-CF-s1 model without second-order collaborative signals performs the worst on each dataset. This result indicates that leveraging the second-order collaborative signals in the graph convolution layer is beneficial to improve the rating prediction accuracy.
- The IGCN-CF-1 model with a linear prediction layer also achieves good prediction performance, which is slightly worse than the basic model. This result indicates that: on the one hand, the graph convolution layer has been able to learn about node embedding representations accurately, and even using simple linear prediction methods can achieve a good prediction performance; on the other hand, using a nonlinear prediction layer can further improve the prediction accuracy of the model.

5.5 Effect of graph convolution layer numbers

Note that NGCF, LightGCN and IGCN-CF stack multiple graph convolutional layers to achieve layer-by-layer refinement of node embeddings. We experimentally examine the

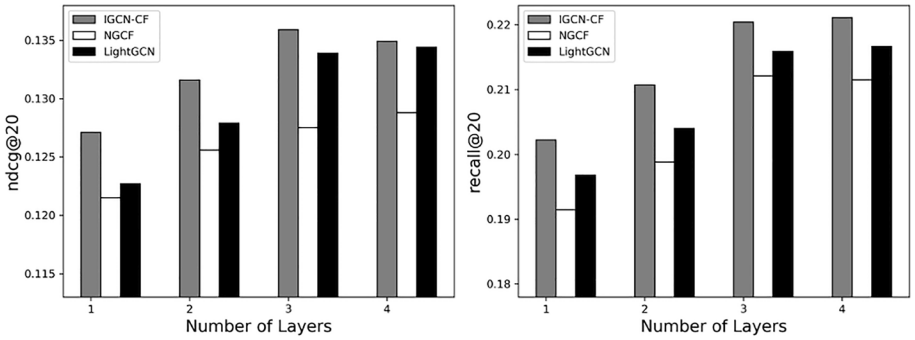


Fig. 8 Results of three models at different layer numbers on Book-crossing

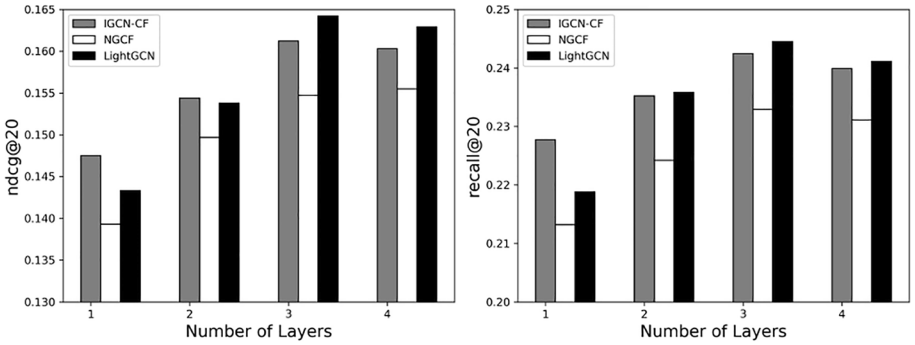


Fig. 9 Results of three models at different layer numbers on Yelp2018

prediction performances of theee models when taking different numbers of graph convolutional layers. Fig. 7, Fig. 8, and Fig. 9 shows the results.

According to the figures, We have the following observations:

- The prediction performances of the three models are obviously affected by the graph convolution layer numbers. They show a similar trend on three datasets: the prediction performances first increase significantly with the increasing of layer numbers, but after 3 layers, they no longer increase significantly and show a slight decrease. This phenomenon indicates that stacking too many graph convolution layers is prone to overfitting problems. Therefore, the maximum number of layers of the model in this paper is set to 3.
- In most cases (except on the Yelp2018 dataset), the prediction performances of our model are significantly better than those of the NGCF and LightGCN model, especially when the number of graph convolutional layers is smaller. This phenomenon is attributed to the advantage of the enhanced graph convolutional layers used in the IGCN-CF model, which improves the accuracy of node embeddings learning by aggregating the second-order collaborative signals inside the layers and incorporating the user's opinions. Thus our model achieves excellent prediction accuracy even with fewer graph convolutional layers.

6 Conclusion

In this paper, we propose an end-to-end collaborative recommendation model based on an improved graph convolutional network for the rating prediction in recommender systems. The model leverages the attribute interactions of entities and uses the aggregated attribute vectors as input, which is more conducive to learning node embeddings. Moreover, it obtains an improved graph convolution layer by adding second-order collaborative signals and incorporating user opinions. The model stacks multiple layers to capture the collaborative signals on the bipartite graph, improving the accuracy of node embeddings learning. Finally, the model uses a nonlinear MLP network to achieve rating prediction. Experimental results on several recommendation datasets show that the prediction accuracy of this model outperforms state-of-the-art recommendation models.

The entity attribute graphs in our model are complete graphs with high memory overhead. The model needs to store the adjacency matrix of the bipartite graph at runtime, which may encounter a storage bottleneck on the vast graph. In future work, we will study the node sampling technique based on the random walk (He & Chua, 2017) to improve the model in this paper so that it can be adapted to very large scale recommendation applications.

Author Contributions Yichen He and Yijun Mao wrote the main manuscript text. Yichen He and Wanrong Gu wrote the key code and conducted the experiments. Xianfen Xie conceived and designed the analysis. All authors reviewed the manuscript.

Funding This work was financially supported by the Philosophy and Social Science Planning Project of Guangdong (GD19CGL34), the Guangdong Computational Science Key Laboratory Development Fund Project (2021010), the National Statistical Science Research Project (2020LY018), the National Social Science Foundation of China (18ZDA093), and Natural Science Foundation of Guangdong (2022A1515011489).

Data Availability Data and Codes are available uponrequest.

Declarations

Ethics approval and consent to participate All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Consent for Publication Authors give full consent for publication.

Competing interests The authors declare that they have no conflict of interest.

References

Kluver, D., Ekstrand, M.D., & Konstan, J.A. (2018). Rating-based collaborative filtering: Algorithms and evaluation. Social Information Access, pp 344–390, https://doi.org/10.1007/978-3-319-90092-6_10.

Jalili, M., Ahmadian, S., Izadi, M., Moradi, P., & Salehi, M. (2018). Evaluating collaborative filtering recommender algorithms: A survey. *IEEE Access*, *6*, 74003–74024. https://doi.org/10.1109/ACCESS.2018.2883742.

Mehta, R., & Rana, K. (2017). A review on matrix factorization techniques in recommender systems. In *2017 2nd International conference on communication systems, computing and it applications (CSCITA), IEEE. pp 269–274*. https://doi.org/10.1109/CSCITA.2017.8066567.

Quadrana, M., Cremonesi, P., & Jannach, D. (2018). Sequence-aware recommender systems. *ACM Computing Surveys (CSUR)*, *51*(4), 1–36. https://doi.org/10.1145/3190616.

Liu, J., & Wu, C. (2017). Deep learning based recommendation: A survey. In K. Kim N. Joukov (Eds.) (eds.) Information *Science and Applications 2017 pp 451–458*. Springer, Singapore. https://doi.org/10.1007/978-981-10-4154-9_52.

Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K.-I., & Jegelka, S. (2018). Representation learning on graphs with jumping knowledge networks. In *International conference on machine learning*, pp 5453–5462, PMLR. https://proceedings.mlr.press/v80/xu18c.html.

Berg, R.V.D., Kipf, T.N., & Welling, M. (2017). Graph convolutional matrix completion. arXiv:1706.02263.

Li, Y., Zhai, C., & Chen, Y. (2014). Exploiting rich user information for one-class collaborative filtering. *Knowledge and Information Systems*, *38* (2), 277–301. https://doi.org/10.1007/s10115-012-0583-9.

Dong, X., Yu, L., Wu, Z., Sun, Y., Yuan, L., & Zhang, F. (2017). A hybrid collaborative filtering model with deep structure for recommender systems. Proceedings of the AAAI Conference on Artificial Intelligence, vol 31(1). https://doi.org/10.1609/aaai.v31i1.10747.

Guo, L., Liang, J., Zhu, Y., Luo, Y., Sun, L., & Zheng, X (2019). Collaborative filtering recommendation based on trust and emotion. *Journal of Intelligent Information Systems*, *53* (1), 113–135. https://doi.org/10.1007/s10844-018-0517-4.

Wang, X., He, X., Nie, L., & Chua, T.-S. (2017). Item silk road: Recommending items from information domains to social users. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval. SIGIR '17*, pp 185–194. Association for Computing Machinery. https://doi.org/10.1145/3077136.3080771.

Wang, X, Wang, D, Xu, C, He, X, Cao, Y, & Chua, T.-S (2019). Explainable reasoning over knowledge graphs for recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 5329–5336. https://doi.org/10.1609/aaai.v33i01.33015329.

Ren, X., Yin, H., Chen, T., Wang, H., Huang, Z., & Zheng, K. (2021). Learning to ask appropriate questions in conversational recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. SIGIR '21, pp 808–817*. Association for Computing Machinery. https://doi.org/10.1145/3404835.3462839.

Hsieh, C.-K., Yang, L., Cui, Y., Lin, T.-Y., Belongie, S., & Estrin, D. (2017). Collaborative metric learning. In *Proceedings of the 26th international conference on world wide Web. WWW '17, pp. 193–201*. International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3038912.3052639.

Liu, T., & He, Z. (2022). Dlir: a deep learning-based initialization recommendation algorithm for trust-aware recommendation. Applied Intelligence, pp 1–12. https://doi.org/10.1007/s10489-021-03039-1.

Fu, M., Qu, H., Yi, Z., Lu, L., & Liu, Y. (2019). A novel deep learning-based collaborative filtering model for recommendation system. *IEEE Transactions on Cybernetics*, *49*(3), 1084–1096. https://doi.org/10.1109/TCYB.2018.2795041.

Guo, H., Tang, R., Ye, Y., Li, Z., & He, X. (2017). DeepFM: A factorization-machine based neural network for ctr prediction. https://doi.org/10.48550/ARXIV.1703.04247.

He, X, & Chua, T.-S. (2017). Neural factorization machines for sparse predictive analytics. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval. SIGIR '17 pp 355–364*. Association for Computing Machinery. https://doi.org/10.1145/3077136.3080777.

Tay, Y, Anh Tuan, L, & Hui, S.C. (2018). Latent relational metric learning via memory-based attention for collaborative ranking. WWW '18, pp. 729–739. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE. https://doi.org/10.1145/3178876.3186154.

Wang, X, He, X, Wang, M, Feng, F., & Chua, T.-S. (2019). Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. SIGIR'19, pp 165–174*. Association for Computing Machinery. https://doi.org/10.1145/3331184.3331267.

He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., & Wang, M. (2020). Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval. SIGIR'20, pp 639–648*. Association for Computing Machinery. https://doi.org/10.1145/3397271.3401063.

Chen, L., Wu, L., Hong, R., Zhang, K., & Wang, M. (2020). Revisiting graph based collaborative filtering: A linear residual graph convolutional network approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(01), 27–34. https://doi.org/10.1609/aaai.v34i01.5330.

Song, W., Xiao, Z., Wang, Y., Charlin, L., Zhang, M., & Tang, J. (2019). Session-based social recommendation via dynamic graph attention networks. In *Proceedings of the twelfth ACM international*

conference on web search and data mining. WSDM '19 pp 555–563. Association for Computing Machinery. https://doi.org/10.1145/3289600.3290989.

Feng, C., Liu, Z., Lin, S., & Quek, T.Q.S. (2019). Attention-based graph convolutional network for recommendation system. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp 7560–7564*. https://doi.org/10.1109/ICASSP.2019.8683050.

Fan, W., Ma, Y., Li, Q., He, Y., Zhao, E., Tang, J., & Yin, D. (2019). Graph neural networks for social recommendation. In *The World wide web conference. WWW '19, pp 417–426*. Association for Computing Machinery. https://doi.org/10.1145/3308558.3313488.

Wang, X., He, X., Cao, Y., Liu, M., & Chua, T.-S. (2019). Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD International conference on knowledge discovery & data mining. KDD '19, pp 950–958*. Association for Computing Machinery. https://doi.org/10.1145/3292500.3330989.

Hui, B., Zhang, L., Zhou, X., Wen, X., & Nian, Y. (2022). Personalized recommendation system based on knowledge embedding and historical behavior. *Applied Intelligence*, *52*(1), 954–966. https://doi.org/10.1007/s10489-021-02363-w.

Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W.L., & Leskovec, J. (2018). Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. KDD '18, pp 974–983*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3219819.3219890.

Wang, J., Huang, P., Zhao, H., Zhang, Z., Zhao, B., & Lee, D.L. (2018). Billion-scale commodity embedding for e-commerce recommendation in alibaba. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. KDD '18 pp 839–848*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3219819.3219869.

Zhang, J., Shi, X., Zhao, S., & King, I. (2019). STAR-GCN: Stacked and Reconstructed Graph Convolutional Networks for Recommender Systemso. https://doi.org/10.48550/ARXIV.1905.13129.

Xiang, R., Neville, J., & Rogati, M. (2010). Modeling relationship strength in online social networks. In *Proceedings of the 19th international conference on world wide Web. WWW '10 pp 981–990*. Association for Computing Machinery https://doi.org/10.1145/1772690.1772790.

Su, Y., Zhang, R.M., Erfani, S., & Gan, J. (2021). Neural graph matching based collaborative filtering. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. SIGIR '21 pp 849–858*. Association for Computing Machinery. https://doi.org/10.1145/3404835.3462833.

Easley, D., & Kleinberg, J. (2010). *Networks, Crowds And Markets: Reasoning About A Highly Connected World*. London: Cambridge University Press.

Wang, J., De Vries, A.P., & Reinders, M.J.T. (2006). Unifying user-based and item-based collaborative filtering approaches by similarity fusion. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval. SIGIR '06 pp 501–508*. Association for Computing Machinery. https://doi.org/10.1145/1148170.1148257.

Bruna, J., Zaremba, W., Szlam, A., & LeCun, Y. (2013). Spectral networks and locally connected networks on graphs. https://doi.org/10.48550/ARXIV.1312.6203.

Kingma, D.P, & Ba, J. (2014). Adam: A method for stochastic optimization. https://doi.org/10.48550/ARXIV.1412.6980.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

◆ 中国最具国际影响力学术期刊
◆ 中文社会科学引文索引（CSSCI）来源期刊
◆ 中国人文社会科学期刊AMI综合评价A刊核心期刊
◆ 中文核心期刊要目总览核心期刊
◆ RCCSE中国核心学术期刊（A）
◆ 中国科技核心期刊
◆ 全国高校百强社科期刊
◆ 全国高校社科名刊

ISSN 1007-3116
CN 61-1421/C

统计与信息论坛

JOURNAL OF STATISTICS AND INFORMATION
TONGJI YU XINXI LUNTAN

2021 6
总第249期 ◎ 第36卷

统计与信息论坛

月 刊
2021 年第 6 期
第36卷 (总第249期)

主管单位
陕西省教育厅
主办单位
西安财经大学
中国统计教育学会高教分会

《统计与信息论坛》编辑委员会

学 术 顾 问 (按姓氏笔画排列)
许宪春 肖红叶 邱 东
胡 健 贺 铿

编委会主任 方 明

编委会副主任 李佐瑞 赵彦云

编委会委员 (按姓氏笔画排列)
万东华 王兆军 王振龙
王浩鸣 方 明 东方社奇
叶植材 史代敏 冯根福
宁泽逵 任 燕 任保平
向寿生 刘 扬 杨 涛
李 元 李金昌 李国武
李佐瑞 李 萍 严明义
吴旺延 张 虎 张宝学
张爱婷 张维群 林共进
金勇进 周 勇 孟晓犁
赵彦云 洪永淼 姚树俊
耿 直 徐勇勇 郭建华
董 箫 蒋 萍 傅德印
谢邦昌

主 编 李佐瑞
副 主 编 姚树俊 张爱婷

目 次

【统计理论与方法】

创新环境指数构建及基于自适应 ARIMA 的动态预警监控研究

谢贤芬,王斌会,古万荣,杨 颖(3)

【经济统计】

人口红利衰减、产业结构调整对中国工业经济发展影响研究

吴 滨,肖 尧(14)

【社会与管理统计】

多维相对贫困测度研究 方迎风,周少驰(21)

低收入群体代际收入流动性变动趋势 徐晓红,曹萍萍(31)

行政区划调整与企业资源配置效率 任宗哲,张陈一轩(41)

基于文本挖掘的数字化水平与运营绩效研究

任碧云,郭 猛(51)

【资源与环境统计】

低碳视角下区域物流与生态环境协调发展研究 周 泰(62)

结构突变下国际油价波动与中国石油行业发展的多重联动关系

吕 政,胡晨沛(73)

【高质量发展专栏】

生产性服务业是推动经济高质量发展的"新引擎"吗?

曹景林,张子贤(84)

中国装备制造业高质量发展的测度及发展路径

钞小静,刘 璐,孙艺鸣(94)

城市创新效率与经济高质量发展的耦合关系及其时空分异

特征研究 杨卫丽,谭景柏,刘道辉,唐 飞(104)

粤港澳大湾区经济高质量发展的评价与靶向路径研究

凌连新,阳国亮(120)

期刊基本参数:CN 61-1421/C * 1986 * m * A4 * 128 * zh * P * ¥15.00 * 1000 * 12 * 2021-06

【统计理论与方法】

创新环境指数构建及基于自适应 ARIMA 的动态预警监控研究

谢贤芬[1a]，王斌会[1b]，古万荣[2]，杨 颖[3]

（1. 暨南大学 a. 经济学院；b. 管理学院，广东 广州 510632；2. 华南农业大学 数学与信息学院，广东 广州 510642；3. 中南大学 法学院，湖南 长沙 410083）

摘要：创新环境问题一直是地区经济发展高度重视的问题，也是学者们对区域创新能力长期研究的重点内容之一。根据高新技术企业的现实情况，结合国内外相关问题的研究成果，从经济因素、制度因素、人才因素、硬件能力及产业聚集五个层面建立了创新环境状态指标体系，采用稳健主成分合成指数法对广东省高新技术企业 1997—2019 年的年度数据构建宏观创新环境指数 MacroIEI，结果显示该方法得到的指数值与实际较为相符。在此基础上，利用自适应 ARIMA 方法对创新环境指数建立预测模型，将模型得到的预测结果加入到改进型 EWMA 控制图中，实现对创新环境指数的动态过程控制。为了验证模型的可靠性，与常规 EWMA 控制图进行过程控制比较研究，研究发现：对创新环境指数 MacroIEI 进行预警监控，改进型 EWMA 控制图具有比常规 EWMA 控制图更好的检测效果。通过创新环境指数 MacroIEI 的过程控制结果分析，证实提出的动态预警监控方法是合理的，政府决策部门或相关企业可以依据这些来评价区域创新环境的状态并作出决策。

关键词：创新环境指数；自适应 ARIMA；预警监控；改进型 EWMA 控制图

中图分类号：F832.1；O212 **文献标志码**：A **文章编号**：1007－3116(2021)06－0003－11

引用格式：谢贤芬，王斌会，古万荣，等. 创新环境指数构建及基于自适应 ARIMA 的动态预警监控研究[J]. 统计与信息论坛，2021，36(6)：3-13.

Citation Form: XIE Xian-fen, WANG Bin-hui, GU Wan-rong, et al. Innovation Environment Indicator Construction and Dynamic Early Warning Monitoring Research Based on Adaptive ARIMA Method [J]. Journal of Statistics and Information, 2021, 36(6): 3-13.

一、引言及文献述评

创新是推动国家和地区经济增长和发展的驱动力，在经济发展过程中，单纯依赖市场的自由引导或企业的自主创新力量往往不够，需要依靠整个创新系统。创新系统是各级政府部门、高等院校、地区科研机构及企业等实体单位协调运作，形成的有机复杂网络系统，在面对全球经济危机及适应城市经济衰退的冲击下，区域或省域经济区的创新系统表现出了较完备的抗风险能力[1]。为此，学术界对创新系统从各方面进行了深入研究，创新理论得到了迅速发展。从地区跨度来看，创新系统分为国家级创新系统、区域级创新系统和城市级创新系统。典型的国家级创新系统包括全球创新指数 GII 及欧洲创新记分牌 EIS，以及中国两个国家创新能力评价系统：中国创新指数和国家创新指数，这些均用来反映国家的创新能力及发展趋势[2]。国内

收稿日期：2020－10－15

基金项目：广东省科技计划项目“广东省创新环境指数构建及预警监测研究——基于高新技术企业数据分析”(2018A070712021)；全国统计科学研究重点项目“多源异构混频大数据模型研究及在经济运行中的应用”(2019LZ37)；全国统计科学研究一般项目“复杂多维数据的时空监控方法研究及其在经济社会中的应用”(2020LY018)

作者简介：谢贤芬，女，广东韶关人，博士，高级实验师，硕士生导师，研究方向：经济统计与预测决策分析；

王斌会，男，陕西西安人，博士，教授，博士生导师，研究方向：稳健统计及预警监控；

古万荣（通讯作者），男，广东韶关人，博士，讲师，硕士生导师，研究方向：数据挖掘算法。

外学者对区域级创新系统及城市级创新系统做了广泛的研究，美国用来评价城市创新程度的指标体系是“创新力指数”，澳大利亚和英国也使用该指数来衡量城市创新程度[3]。中国区域级或省级创新系统中，较为认可的有《中国区域创新能力报告》、中国城市创新力指数和中国创新能力评价。纵观国内外学者的研究，衡量创新系统的各类研究也日益增多，主要有：创新能力研究、创新环境研究、创新发展研究、创新绩效研究、创新空间关联研究等[4-6]。

本文主要研究创新系统中的创新环境。20 世纪 80 年代 GREMI 提出创新环境的概念，最初的涵义主要包括创新环境的定义及构建、创新环境与经济发展的关系、创新环境对创新绩效的影响研究三个方面。创新环境是指某一特定区域内聚集资源的再分配，包含两个层面的涵义：第一个层面是指单个或少数几个关联企业内部的创新环境，如企业管理者的冒险、企业应对市场机会的能力、企业自身资源的合理配置等，这是微观层面的创新环境状态；第二个层面是指企业所处的外部环境创新，包括经济环境、劳动力市场、技术水平、区域文化等因素，这是宏观层面的创新环境状态。

创新环境状态及指数研究最近受到国内外学者的重视，由于不同文献研究的数据、时间范畴、模型、国情等因素的不同，分析过程和得出的结论也不尽相同。国内外相关研究多见于创新环境指标体系研究与构建、创新环境的测度方法、创新环境发展的对策研究等方向。在创新环境指标体系构建的研究工作中，国外学者 Michael E. Porter 等基于 17 个区域 20 多年的数据进行研究，发现硬件条件完善有利于提升区域的创新环境[7]；Porter 认为产业聚集有利于良好创新环境的创造，产业集群内的企业往往能够更快发展，以更低的成本快速实现企业的创新，同时，相关政策制度及市场也是影响创新环境的重要因素[8]；学者 Martina Fromhold-Eisebith 从社会资本对创新环境的影响因素分析出发，结合德国 Aachen 地区的真实情况，分析两者之间的相关性及差异性，发现相关性及差异性是相互作用的，这种协同作用有利于提升创新环境[9]。Collins 等研究了欧洲企业创新的政策环境，发现地区的政策环境能够极大地鼓励并引导企业进行技术创新，对创新活动推进具有显著的意义[10]；杨毅同样认可政策环境的重要性，认为应加大政策层面对创新环境的引导和扶持力度，从而提升技术创新活动的活跃程度[11]；施蕾等从基础条件、人力资源、孵化能力和研发转移 4 个层次构建了包括 20 个基础指标的创新环境评价指标体系，从省域的角度进行比较研究，结合陕西省实际情况，深入探究其不足之处，并给出相应的对策建议[12]。李玲等首次构建了区域包容性创新环境指标体系，该指标体系既涵盖了创新环境的基础设施、支撑条件、市场及金融环境方面，也涉及创新主体、创新投资、企业的创新能力和制度环境等四个方面[13]。

在创新环境测度分析方面，大部分研究主要是建立各种模型评价各省份在一个时间点的创新环境，分析区域的创新能力差距并尝试解释产生差距的原因，它们并不能对各区域的创新环境状态进行纵向的动态分析。崔航采用层次分析法对中国典型的六大城市群（长三角、珠三角、京津翼、成渝、武汉、哈长）进行实证分析，基于收集的创新环境指标数据进行评价，对各个城市群的不同维度进行了详细分析，同样采用层次分析法的有党晶晶等人[14-15]。黄玲雁构建涵盖经济环境、金融环境、基础设施环境因素的创新环境指标体系，利用因子分析法对湖南省 14 个地级市的创新环境进行评价与分析，揭示了省域间的差异性[16]。陈红川采用 CURE 聚类分析算法对高新技术企业的创新环境进行了综合分析，从政策环境、市场环境、人才环境等多个角度进行了细化及分析[17]。鉴于多数文献常采用单一的评价方法进行测度，李沃源等将层次分析法、集对分析法、灰色关联法和 TOPSIS 结合起来，建立了基于离差最大化的组合评价模型，该模型解决了应用单一方法的不足，并对西部地区科技创新环境的区域差异性进行分析[18]。另外还有一些文献是以创新环境来解释区域创新能力或区域创新效率，结合各种模型研究创新环境对它们的影响[19]。

创新环境发展的对策研究方面，苏炜从基础设施环境、经济环境及创新的产出等多个方面综合评价了珠三角九大国家级自主创新示范区的创新环境，结合各地的实际情况，提出了持续高创新投入、加强基础设施建设工作、建立多元化的创新投入新机制、完善创新体制等多项政策措施[20]。林媛媛等通过分析福建省的环境规划、企业技术创新现状和环境规制低效率原因三方面因素，对福建省企业的发展提供了研究对策[21]。

纵观学者们的研究，他们大多采用的是事件分析法，即在构建创新环境指标体系的基础上，分析过去确定时间段的创新环境状态，但是缺乏对未来创新环境发展状态走向的预测及监控。如何衡量一个国家或地区的创新环境状态并进行预测呢？笔者认为应构建一个综合性指标（指数）来进行测度，目前暂未发现有其

他文献构建创新环境指数，对创新环境指数进行预警监测的相关文献也未见报道。

本文从宏观角度去研究创新环境状态的相关因素，构建指标体系，进而建立预警机制，而不是纯粹作为区域创新能力或区域创新效率的解释变量。另外，创新环境发展过程涉及的经济运行指标之间往往存在着相关关系，传统的主成分合成指数法通常要求数据相互独立，直接采用传统主成分合成指数法可能得到不合理的结论。因此，这里利用稳健主成分合成指数法构建宏观创新环境指数 MacroIEI（Macro Innovation Environment Indicator），基于自适应 ARIMA 方法及改进型 EWMA 控制图建立创新环境指数的动态预警模型，研究指标的动态规律并给出相应的政策建议，以供决策部门或相关企业参考，这为中国区域创新环境的测度提供了新的研究思路，对创新环境预警系统的建立也有一定的意义[22]。

二、动态预警模型与方法

（一）创新环境状态指标体系

本文在宏观层次下构建创新环境状态指标体系，指标选取应能覆盖影响创新环境的经济与社会因素、硬件投入、市场机会等方面。这里参考了国家统计局社科文司“中国创新指数（CII）研究”课题组（2014）提出的指标体系框架，该文提出创新环境领域应能反映驱动创新能力发展必备的人力、财力等基础条件的支撑情况，以及政策环境对创新的引导和扶持力度，共包括经济活动人口中大专及以上学历人数、人均 GDP、信息化、科技拨款占财政拨款的比重、享受加计扣除减免税企业所占比重五个指标[23]。然而在实际应用中，有些参数可能会缺失，可以用相关或相近指标来代替。比如，这里的信息化及享受加计扣除减免税企业所占比重两个指标数据缺失，同时使用本科以上学历人数指标替代大专及以上学历人数指标、产值超亿元企业数量替代享受加计扣除减免税企业等。

鉴于上述的数据可获取性以及统计口径的差异性，从经济因素、制度因素、人才因素、硬件能力及产业聚集五个层次出发，选取的创新环境状态指标主要包括专业机构服务数、入园企业数量、企业固定资产投资、财政科技支出、高新区鼓励创新的政策及措施数量、本科以上学历人数、R&D 人员数、高新区从业人员数、R&D 经费支出、高新区高新技术产品出口总额、高新区 GDP、高新技术产品产值、产值超亿元企业数量等 13 个相关指标。有些原始数据不能很好地反映创新环境状态，本文提出构建比率指标，即用增长率＝（当期指标值－上一期指标值）/上一期指标值来表示，总体创新环境指标体系见表 1。

表 1　创新环境状态指标体系

层次	指标
经济因素	专业机构服务数增长率（X_1）
	入园企业数量增长率（X_2）
	财政科技支出增长率（X_3）
	企业固定资产投资增长率（X_4）
制度因素	高新区相关政策措施数量增长率（X_5）
人才因素	本科以上学历占高新区从业人员比重（X_6）
	R&D 人员占高新区从业人员比重（X_7）
硬件能力	R&D 经费支出增长率（X_8）
	高新技术产品出口总额增长率（X_9）
产业聚集	高新区 GDP 增长率（X_{10}）
	高新技术产品产值增长率（X_{11}）
	产值超亿元企业数量增长率（X_{12}）

专业机构服务数增长率及入园企业数增长率用来反映地区经济的发展能力；财政科技支出增长率及企业固定资产投资增长率用来反映地区的创新投入及市场经济水平；高新区相关政策措施增长率指标表明国家/地区对创新园区从宏观角度进行引导及支持的程度，是政府政策调控的有力支撑；本科以上学历占高新区从业人员比重及 R&D 人员占高新区从业人员比重，能反映地区的劳动者结构及整体素质，是创新发展必备的人力条件；R&D 经费支出增长率和高新技术产品出口总额增长率，反映一个地区创新环境的研发投入及硬件水平；高新区 GDP 增长率、高新技术产品产值增长率和产值超亿元企业数量增长率三大指标，是地区产业聚集效应的最直接表现，而产业聚集跟产出规模又是相互影响的，较大的产出规模将吸引更多优势企业聚集，产业的聚集又提高了资源共享，提高地区竞争力，从而实现更大的产出规模。

（二）创新环境指数计算方法

主成分分析方法的全称为 Principal Component Analysis，通常简记为 PCA，该方法最早是由 Pearson 在 1901 年提出来的，现在使用的 PCA 技术是经过 Hotelling 改进的，其基本思想便是数据降维，通过对数据压缩把原来若干个映射同一个问题的研究数据 x_i 解释成一个新的相互无关的综合指标 y_i，这个指标被称为

主成分，它们之间有如下的公式：

$$\begin{cases} y_1 = u_{11}x_1 + u_{12}x_2 + \cdots + u_{1p}x_p = u'_1 x \\ y_2 = u_{21}x_1 + u_{22}x_2 + \cdots + u_{2p}x_p = u'_2 x \\ \cdots \\ y_p = u_{p1}x_1 + u_{p2}x_2 + \cdots + u_{pp}x_p = u'_p x \end{cases} \tag{1}$$

记标准化数据矩阵为 $Z_{n\times p}$，建立变量的相关系数矩阵：$R=(r_{ij})_{p\times p}=Z'Z$，然后求出相关矩阵 R 的特征值 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p > 0$ 及相应的单位特征向量 $u_1, u_2, \cdots, u_p$，写出主成分：$y_i = u_{i1}z_1 + u_{i2}z_2 + \cdots u_{ip}z_p$，这里 $i=1,2,\cdots,p$。

根据累积方差贡献率高于 80%的基本原则，选定前 m 个主成分反映原评价对象。设前 m 个主成分为：$C_1, C_2, C_3, \cdots, C_m$，以方差贡献率为权重，可以得到综合评价函数：

$$C = \frac{\lambda_1 c_1 + \lambda_2 c_2 + \cdots + \lambda_m c_m}{\lambda_1 + \lambda_2 + \cdots + \lambda_m} = \sum_{i-1}^{m} w_i C_i \tag{2}$$

将主成分综合得分值作为主成分合成指数值(PCSI)，PCSI=C。通过该指数可以很好地反映宏观经济运行指数的动态变化过程。

然而，从上面的传统主成分构建过程可以看到，主成分的计算基于协方差矩阵(或标准化矩阵)，并在此基础上计算其他统计量。由于协方差矩阵是不稳健的，因此样本中出现离群值时，相关系数矩阵、特征值和特征向量会受到较大影响，进而得到不准确的计算结果。2007 年王斌会提出了基于 FAST-MCD 算法构建稳健的均值向量和协方差矩阵，并应用到主成分分析中，通过模拟和实证研究发现，改进后的主成分分析方法确实能够对离群值起到很好的抵抗作用[22]。稳健主成分合成指数构建的具体算法如下：

(1)获取经济运行样本数据，并进行标准化，记为 Z；

(2)基于 Fast-MCD 算法得到协方差矩阵 S 和稳健均值向量 T；

(3)通过协方差矩阵 S 得到稳健相关矩阵 R；

(4)求解矩阵的特征方程，可得稳健相关矩阵 R 的特征值 λ_p 和特征向量 u_p；

(5)写出主成分 $y_i(i=1,2,\cdots,p)$，按累积方差贡献率大于 80%的原则，取前 m 个主成分计算综合得分。按上述方法得到的综合得分值即为稳健主成分合成指数值，记为：PCSI_m。本文将以此构建宏观创新环境指数 MacroIEI，即：MacroIEI=PCSI_m。

(三)基于自适应 ARIMA 方法的创新环境指数预警模型

1. 基于自适应 ARIMA 的预测方法

ARIMA 模型(自回归求积移动平均模型)是一个经典的时间序列预测方法，其基本原理是：先利用 d 阶差分将非平稳时间序列平稳化，然后对平稳化后的时间序列，利用自回归过程 AR(p)、移动平均过程MA(q)以及样本自相关系数(ACF)和偏自相关系数(PCF)等对模型进行辨识，该模型常用于时间序列分析。

如果某个时间序列$\{X_t\}$的 d 次差分 $W_t=(1-B)^d X_t$ 是一个平稳的 ARMA 过程，则称$\{X_t\}$为自回归移动平均求和模型。如果 W_t 服从 ARMA(p,q)模型，则称$\{X_t\}$是 ARIMA(p,d,q)过程。ARIMA 包含了自回归、差分以及移动平均三个组成部分，p、d、q 分别表示自回归阶数、差分次数和移动平均的阶数。在实际应用中，通常取 $d=1$ 或 2。这时可用平稳过程 W_t 来代替不平稳的 X_t 在 ARMA 模型中的位置，即：

$$W_t = c + \phi_1 W_{t-1} + \cdots + \phi_p W_{t-p} + \varepsilon_t + \theta_t \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \tag{3}$$

用滞后算子表示，则有：

$$\Phi(B)W_t = c + \Theta(B)\varepsilon_t \tag{4}$$

其中，$\Phi(B)=1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p$，$\Theta(B)=1+\theta_1 B+\theta_2 B^2+\cdots+\theta_q B^q$。经过 d 阶差分变换后的 ARMA(p,q)模型称为 ARIMA(p,d,q)模型，式(3)等价于下式：

$$\Phi(B)(1-B)^d X_t = c + \Theta(B)\varepsilon_t \tag{5}$$

其中$\{\varepsilon_t\}$是白噪声过程。模型 ARIMA(p,d,q)的识别和参数估计步骤与 ARMA(p,q)模型的参数估计基本类似。对于 ARIMA(p,d,q)模型，参数 d 的确定由差分次数来决定。ARMA(p,q)模型的识别方法较多，一般会利用 ACF 函数和 PACF 函数来判断类型。初始值的选择往往是不准确的，需要根据模型的检验结

果对参数不断进行调整，最终得到最优模型。估计模型的参数有两种常见的方法：最小二乘法和极大似然法，这两种方法都是“精估计”。本文提出一种自适应方法，利用 R 统计软件自编了 adaptive. Arima()函数，对 R 自带的 auto. arima()函数进行了改进。该函数能够绘制模型的识别过程图，并给出最优模型的参数估计值及序列的预测值。另外，由于经济运行数据可能具有季节因素，如 CPI 指数的周期长度为 12，这里在实现 adaptive. Arima()时考虑了对数据进行季节 ARIMA 模型的估计。adaptive. Arima()的基本模型如下：

adaptive. Arima (y, d = NA, D = NA, max. p = 5, max. q = 5, max. P = 2, max. Q = 2, max. order = 5, max. d = 2, max. D = 1, …)

adaptive. Arima()的基本思想是对 ARIMA 建模的系数进行逐步向后选择的过程，其原理类似于线性回归模型的逐步回归过程，基于极大似然估计和 AIC/BIC/AICC 进行模型的选择，这里主要采用 AIC 最小准则选择最优模型。

图 1 是以高新区 GDP 指数为例说明 adaptive. Arima()的模型识别过程，每一行对应于每一步的后向选择。每个小方格中的数字是对应的系数，系数值越大其对应的小方格颜色就越深。每个小方格右下角的小三角形代表该系数的 p 值，p 值介于 0～0.1 之间，p 值越大，对应的灰色就越深，p 为 0.1 时为黑色。图形的左边是在每一步骤中建立的模型的 AIC 值，右侧的 sar 是对应的 AR 模型的残差项系数值，sma 是对应的 MA 模型的残差项系数值。根据 AIC 最小准则，这里选择迭代 4 次之后建立的 ARIMA 模型。从图 1 及 adaptive. Arima()函数的运行结果得到的最优模型为：

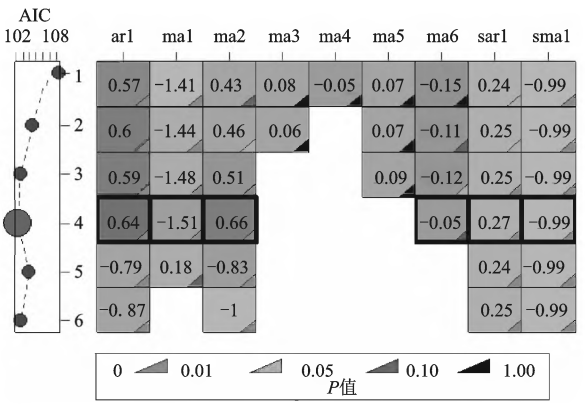


图 1　1997—2019 年高新区 GDP 的 ARIMA 模型识别过程图

$$\text{ARIMA}(p=1,d=1,q=c(1,2,6))\times(P=1,D=1,Q=1)_7$$

2. 构建基于改进型 EWMA 的动态预警模型

1959 年 Roberts 提出了指数加权移动平均(EWMA)控制图，它与累积和控制图(CUSUM)一样，要求观测序列值相互独立。然而，经济运行时间序列数据往往存在着自相关现象，若直接采用传统的 EWMA 控制图进行过程控制，估计出的过程方差是有偏的。这里，将采用考虑了自相关关系的改进型 EWMA 控制图，并把上节中基于自适应 ARIMA 方法得到的预测结果加入到该控制图中，从而实现对创新环境指数进行动态过程控制。改进型 EWMA 控制图的原理如下：

设$\{X_t\}$是离散平稳过程，EWMA 统计量 Z_t、自协方差函数 $R(\tau)=\text{cov}(X_t,X_{t+\tau})=E[(X_t-\mu)(X_{t+\tau}-\mu)]$仅与滞后期有关。统计量 Z_t 及序列$\{Z_t\}$的自协方差函数如下：

$$Z_t=\lambda x_t+(1-\lambda)Z_{t-1}=\lambda\sum_{i=0}^{t-1}(1-\lambda)^i X_{t-i}+(1-\lambda)^i Z_0 \tag{6}$$

$$\begin{aligned}\text{cov}(Z_t,Z_{t+\tau})&=\text{cov}\Big[\lambda\sum_{i=0}^{t-1}(1-\lambda)^i X_{t-i},\lambda\sum_{j=0}^{t+\tau-1}(1-\lambda)^j X_{t+\tau-j}\Big]=\lambda^2\sum_{i=0}^{t-1}\sum_{j=0}^{t+\tau-1}(1-\lambda)^{i+j}R(j-i-\tau)\\&=\lambda^2\Big\{\sum_{k=-(t-1)}^{0}\sum_{i=-k}^{t-1}(1-\lambda)^{2i+k}R(k-\tau)+\sum_{k=1}^{\tau}\sum_{i=0}^{t-1}(1-\lambda)^{2i+k}R(k-\tau)+\\&\quad\sum_{k=\tau+1}^{t-1+\tau}\sum_{i=0}^{t-1-\tau-k}(1-\lambda)^{2t-k}R(k-\tau)\Big\}\\&=[\lambda/(2-\lambda)]\sigma_x^2\times\Big\{\sum_{k=0}^{t-1}\rho(\tau+k)(1-\lambda)^k[1-(1-\lambda)^{2(t-k)}]+\\&\quad\sum_{k=1}^{\tau}\rho(\tau-k)(1-\lambda)^k[1-(1-\lambda)^{2t}]+\sum_{k=\tau+1}^{t-1+\tau}\rho(\tau-k)(1-\lambda)^k\times[1-(1-\lambda)^{2(t+\tau-k)}]\Big\}\end{aligned} \tag{7}$$

在式(6)中，λ 是一个常数，取值范围为：$0<\lambda\leqslant 1$，EWMA 统计量的初值 Z_0 一般取 $E(X)=\mu$。在式(7)中，

$\rho(k)=R(k)/R(0)$表示$\{X_t\}$后面k期的自相关系数。当$\tau=0$时，式(7)即为Z_t的方差：

$$D(Z)=\sigma_z^2=[\lambda/(2-\lambda)]\sigma_x^2\times\left\{1-(1-\lambda)^{2t}+2\sum_{k-1}^{t-1}\rho(k)(1-\lambda)^k\times[1-(1-\lambda)^{2(t-k)}]\right\} \tag{8}$$

当$t\to\infty$时，$\mathrm{cov}(Z_t,Z_{t+\tau})$收敛且收敛值不依赖于$t$。由于$\mathrm{cov}(Z_t,Z_{t+\tau})$收敛，则当$t$很大时，由式(7)可知存在常数$M$，使得：

$$\mathrm{cov}(Z_t,Z_{t+\tau})\approx[\lambda/(2-\lambda)\sigma_x^2\times\{\sum_{k-0}^{M}\rho(\tau+k)(1-\lambda)^k[1-(1-\lambda)^{2(M-k)}]+\sum_{k-1}^{\tau}\rho(\tau-k)(1-\lambda)^k[1-(1-\lambda)^{2M}]+\sum_{K-\tau+1}^{M+\tau}\rho(\tau-k)(1-\lambda)^k\times[1-(1-\lambda)^{2(M+\tau-k)}]\}] \tag{9}$$

那么，Z_t的近似方差为：

$$D(Z)=\sigma_z^2\approx[\lambda/(2-\lambda)]\sigma_x^2\times\{1+2\sum_{k-1}^{M}\rho(k)(1-\lambda)^k\times[1-(1-\lambda)^{2(M-k)}]\} \tag{10}$$

而 Hunter 及 Zhang 等人已经证明了 Z 在自相关条件下的数学期望为[24-25]：

$$E(Z)=E(X)=\mu \tag{11}$$

此时，可对 EWMA 控制统计量 Z_t 进行控制，建立的改进型 EWMA 控制图的控制线为：$E(Z)\pm k\sqrt{D(z)}$。一般情况下取 $k=3$，$E(Z)$由式(11)计算得到；$\sqrt{D(z)}$的取值依赖于 t，当 t 很大时利用式(10)计算得到，当 t 较小时使用式(8)来计算，再将 $E(Z)$及$\sqrt{D(z)}$代入控制线的计算公式，即有：

$$E(Z)\pm k\sqrt{D(z)}=\mu\pm 3\sigma_z \tag{12}$$

本文将利用上述方法对创新环境指数 MacroIEI 进行过程控制，若计算得到的统计量观测值全部落在式(12)对应的控制线内，表明不存在离群值；否则，则存在离群值，需对离群值作进一步分析。由于在对创新环境指数进行实际的过程控制时，控制统计量 Z_t 的均值及方差会受到样本均值、样本自协方差及相关系数的影响，为了减少相关因素对控制参数估计准确度的影响，样本数量不宜太少。

三、分析过程和结果

(一)样本数据及描述性统计分析

我们使用创新环境状态指标体系(见表 1)中的广东省高新技术企业年度数据，样本期为 1997—2019 年，指标数据来自《广东统计年鉴》及《广东科技年鉴》，表 2 罗列了各指标变量的定义及基本描述性统计结果。

表 2　描述性统计分析

变量	定义	均值	标准差	中位数	偏度	峰度	变异系数
X_1	专业机构服务数增长率	0.093	0.207	0.127	0.56	3.11	2.219
X_2	入园企业数量增长率	0.050	0.141	0.064	0.90	5.00	2.836
X_3	财政科技支出增长率	0.051	0.089	0.030	0.64	2.64	1.732
X_4	企业固定资产投资增长率	0.048	0.084	0.063	−0.15	1.90	1.754
X_5	高新区相关政策措施增长率	0.065	0.190	0.043	1.10	3.50	2.935
X_6	本科以上学历占高新区从业人员比重	0.473	0.053	0.470	0.49	2.70	0.111
X_7	R&D 人员占高新区从业人员比重	0.312	0.070	0.302	0.60	2.44	0.225
X_8	R&D 经费支出增长率	0.222	0.328	0.139	1.90	6.90	1.475
X_9	高新技术产品出口总额增长率	0.109	0.177	0.134	−0.20	2.61	1.618
X_{10}	高新区 GDP 增长率	0.061	0.123	0.028	1.00	3.60	2.009
X_{11}	高新技术产品产值增长率	0.066	0.150	0.075	0.001	2.00	2.256
X_{12}	产值超亿元企业数量增长率	0.051	0.146	0.059	1.10	5.30	2.860

从表 2 可以看出，只有 X_6(本科以上学历占高新区从业人员比重)及 X_7(R&D 人员占高新区从业人员比重)指标的均值大于标准差，其他变量的标准差均很大，偏离均值较多，说明这些变量是非正态的。比如入

园企业数量增长率 X_2，其均值为 0.050，标准差为 0.141，约为均值的 3 倍，产值超亿元企业数量增长率 X_{12} 的均值为 0.051，标准差为 0.146，也约为均值的 3 倍，这两者都不符合非正态分布。此外，从 Skewness 和 Kurtosis 两个统计量值也可以分析出这些数据不服从正态分布。

这里还引入了变异系数对各指标进行测度，变异系数值越小，说明数据的波动程度越小，变异系数值越大说明数据的波动程度越大，更可能存在离群值，从而影响创新环境的状态。如 X_2 的变异系数为 2.836，X_6 的变异系数为 0.111，说明 X_2 的变动要比 X_6 的大，表示在样本期内入园企业数量增长率变量的波动较大，这与广东省对高新区的投入息息相关，包括财政科技投入、政策环境对企业创新的扶持力度（如企业享受减免税收优惠政策）、产业聚集等因素。另外，衡量地区产业聚集效应的三大指标为：高新区 GDP 增长率（X_{10}）、高新技术产品产值增长率（X_{11}）和产值超亿元企业数量增长率（X_{12}），它们的变异系数值也较大，均大于 2.0，进一步验证了产业聚集和产出规模的正向反馈作用，产出规模提升有利于促进产业聚集，而产业聚集反过来又能实现更大规模产出。

（二）广东省创新环境状态综合评价

基于表 1 的创新环境状态各项指标，运用稳健主成分合成指数法计算得到相关数据见表 3。从表 3 可知，前 5 个特征值的累积方差贡献率超过了 80%，说明这 5 个主成分包含了大部分状态指标的信息，这些主成分对应的载荷矩阵如表 4 所示。

表 3　特征值、方差贡献率及累积贡献率

主成分	特征值	方差贡献率	累积方差贡献率
Comp. 1	1.746 0	0.254 0	0.254 0
Comp. 2	1.474 7	0.181 2	0.435 2
Comp. 3	1.394 8	0.162 1	0.597 3
Comp. 4	1.075 5	0.106 4	0.703 7
Comp. 5	1.056 5	0.098 1	0.801 8
Comp. 6	0.959 1	0.076 7	0.863 4
Comp. 7	0.718 4	0.043 0	0.906 4
Comp. 8	0.635 3	0.033 6	0.940 1
Comp. 9	0.540 0	0.024 3	0.964 3
Comp. 10	0.476 2	0.018 9	0.983 2
Comp. 11	0.344 7	0.009 9	0.993 1
Comp. 12	0.286 8	0.006 9	1.000 0

表 4　主成分载荷

变量	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5
X_1	0.270	0.172	0.251	0.000	0.350
X_2	0.371	−0.236	0.000	0.387	−0.365
X_3	−0.391	0.186	0.296	0.301	0.000
X_4	0.000	0.342	0.465	0.339	0.000
X_5	0.173	0.463	0.000	−0.109	−0.454
X_6	0.000	0.422	−0.445	0.135	0.191
X_7	0.000	0.433	−0.468	0.000	0.164
X_8	0.283	0.000	0.102	0.188	0.554
X_9	0.447	0.000	−0.123	−0.159	0.000
X_{10}	0.424	−0.128	0.000	0.470	0.000
X_{11}	0.165	0.403	0.335	−0.138	−0.361
X_{12}	0.339	0.000	0.254	−0.560	0.159

对前 5 个主成分按方差贡献率进行加权计算，利用创新环境指数计算方法获得各年份的稳健主成分合成指数值，即创新环境指数 MacroIEI，绘制其走势图如图 2 所示。

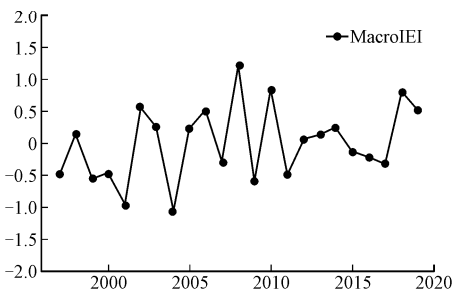


图 2　1997—2019 年广东省创新环境指数走势图

从图 2 可以看出，在整个研究期内，广东省创新环境指数处于一个波动的状态。其中，有 3 个时间段的波动幅度较大，分别为：2001—2002 年、2003—2005 年和 2007—2011 年。以 2008 年至 2009 年为例，2008 年的指数值为1.219 2，2009 年的指数值为−0.589 1，波动差为1.808 3，是整个研究期内波动幅度最大的。总体而言，创新环境指数的变动受到宏观政策、外部环境、地区政府投入、经济因素等方面的影响，后续的预警监控分析中将作更详细的分析。

（三）估计方法和结果

1. 平稳性检验

平稳是建立时间序列模型的前提，为了确定创新环境指数不存在随机趋势或确定趋势，应对该变量进行平稳性检验。本文采用 ADF（Augmented Dickey-Fuller Test）方法来检验该变量序列的平稳性情况。对广东省高新技术产业创新环境指数 MacroIEI 进行平稳性检验，结果显示其不平稳。从图 2 的 MacroIEI 走势

图可以看出，该时间序列不呈现指数增长趋势，因此采用差分法进行平稳化。经一阶差分后，序列 d. MacroIEI 的 ADF 检验结果为平稳，见表 5。

表 5　平稳性检验结果

序列	Dickey. Fuller	P 值	结果
原始序列	−3.1	0.100	不平稳
一阶差分	−4.8	0.010	平稳

注：表中数据是显著性水平为 0.05 时的检验结果。

2. 初步相关性分析及模型确定

对一阶差分 d. MacroIEI 序列进行数据的相关性及正态性检验，从序列数据的 ACF 图和偏 PACF 图（图 3）可以看出，ACF 很明显是“2 后截尾”，而 PACF 图中有若干项非 0 而呈现出振荡拖尾的现象，可初步判断序列 d. MacroIEI 序列应为 ARIMA(p,0,q)模型(p 不为 0)。

另外，对序列数据 d. MacroIEI 进行正态性检验，从其直方图及 QQ 图（见图 4）可以看出，基本符合正态分布。由于数据序列样本量为 20 多个，这里采用小样本的正态性检验方法 Shapiro-Wilk 进行检验，得到的 $W=0.934\ 8$，p 值为0.170 8>0.05，结果认为该样本数据符合正态分布。

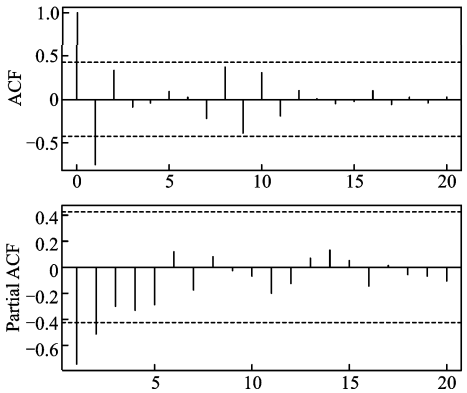


图 3　d. MacroIEI 的自相关图和偏自相关图

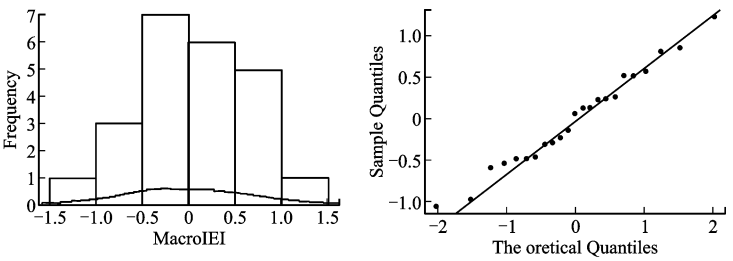


图 4　d. MacroIEI 的直方图和 QQ 图

我们利用提出的自适应 ARIMA 方法，即利用 adaptive. Arima()函数对模型进行识别，并对模型的最佳参数进行估计，以 AIC(Akaike information criterion)最小准则 $AIC=-2\ln(L)+2k$ 和 BIC(Bayesian Information Criterion)最小准则 $BIC=-2\ln(L)+k\ln(n)$ 来选择最佳模型，其中 L 为最大似然函数，n 为样本数量，k 为模型的参数个数。这里基于实际数据，设置初始值为 $p=6$、$d=1$、$q=1$，初始的完全模型为 ARIMA($p=6$,$d=1$,$q=1$)，即有：

$$Y_t=-1.675Y_{t-1}-1.580Y_{t-2}-1.183Y_{t-3}-0.603Y_{t-4}+0.003Y_{t-5}+0.201Y_{t-6}+0.519\varepsilon_{t-1}$$

该初始模型对应的 AIC 值为 53.93，BIC 值为 55.72，极大似然估计值为 −18.96。基于该初始模型调用 adaptive. Arima()方法进行逐步向后选择，最终得到的最佳模型是：ARIMA($p=4$,$d=1$,$q=0$)，对应的 AIC 值为 48.58，BIC 值为 54.04，极大似然估计值为 −19.92，该模型的系数估计值见表 6。

表 6　创新环境指数差分序列 d. MacroIEI 拟合结果

模型	ar1	ar2	ar3	ar4
系数值	−1.182	−1.017	−0.730	−0.348
s. e.	0.209 2	0.326 5	0.319 3	0.210 2
t 值	−5.649	−3.116	−2.286	−1.655
p 值	1.61E−08	1.84E−03	2.23E−02	9.8E−02

ARIMA(4,1,0)模型为：

$$Y_t=-1.182Y_{t-1}-1.017Y_{t-2}-0.73Y_{t-3}-0.348Y_{t-4}$$

这里，Y_t 为原始序列作 $d=1$ 阶差分后得到的序列，即：d. MacroIEI。此时，基于拟合的 ARIMA 模型，向前预测两期，得到序列 d. MacroIEI 的预测值为0.697 3及0.337 1，将它们代入到一阶差分计算公式，即可得到广东省创新环境指数 MacroIEI 在 2020 年和 2021 年的预测值，分别为−0.184 9和0.152 2，其预测走势图见图 5。

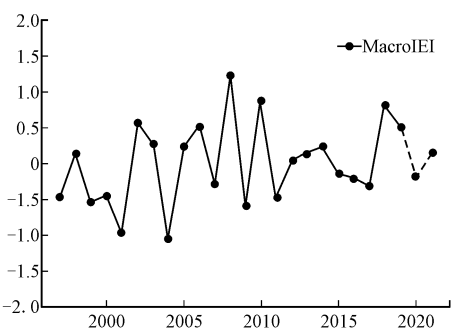


图 5　创新环境指数 MacroIEI 的预测走势图

从图 5 可以看出，2018 年的创新环境指数值比 2017 年的

指数值有较大幅度增长，2019 年的指数值开始下降，比 2018 年下降了0.292 3，下降的幅度较小。预测的 2020 年指数值接着下降，比 2019 年下降了0.697 3，而 2021 年相比 2020 年又是增加的，从－0.184 9增加到 0.152 2，增幅为0.337 1。从整体上看，创新环境指数呈现出缓慢增长的趋势。

3. 建立动态 EWMA 预警监测模型

对加入了预测值的创新环境指数 MacroIEI 数据进行动态过程监控，而使用常规控制图的基本假设前提是：一是质量特征值相互独立；二是质量特征值服从正态分布。实际的经济运行过程中，直接获得的经济运行时间序列数据往往并不满足“服从正态分布”和“数据间互不相关”这两个基本条件。因此，需要对指数 MacroIEI 进行平稳性、正态性及相关性检验。

检验过程类似于上一节，对加入了两期预测数据的 MacroIEI 进行一阶差分，得到序列 d. MacroIEI_p，对该序列进行 ADF 检验，检验结果为 Dickey-Fuller＝－3.92，p 值为 0.00，结果平稳，表明一阶差分后的序列基本消除了趋势。对该序列进行 Shapiro-Wilk 正态性检验，W＝0.978 2，p 值为0.270 3＞0.05，该结果认为样本数据是来自正态分布的总体。这里省略对序列数据作 ACF 图和偏 PACF 图，因为结果跟上文的 d. MacroICI 类似。

由于创新环境指数的一阶差分序列 d. MacroIEI_p 存在自相关现象，这里采用改进型 EWMA 控制图进行预警监测，同时也应用常规 EWMA 控制图进行过程监控，结果分别见图 6、图 7。

在过程控制图中有三条特殊的水平直线，如图 6，UCL 为上控制线，LCL 为下控制线，中间水平直线为中心线 CL。当检测统计量的观测值落在上下控制线外的区域时，认为过程不处于受控状态，控制图将给出过程失控的警报，即该观测值为可能“离群值”。

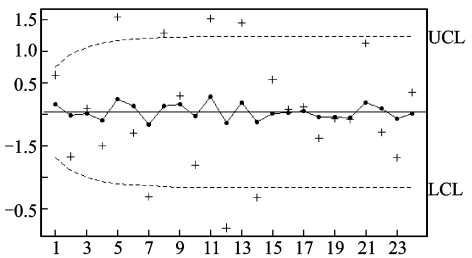


图 6　基于常规 EWMA 控制图的过程控制图

从常规 EWMA 控制图可以看到，所有样本数据点及预测数据均落在控制限区域内，均不存在异常情况。对于改进型 EWMA 控制图，我们可以看到，一共有四个点是落在控制线外的，分别是 2002 年、2008 年、2009 年、2010 年，而其他年份的真实值及两个预测数据值都落在控制线内。

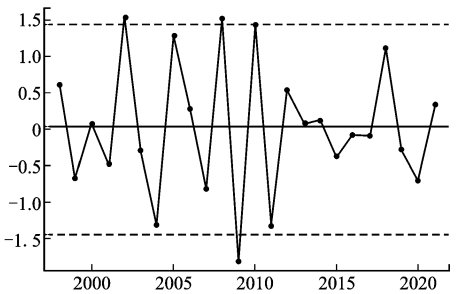


图 7　基于改进型 EWMA 控制图过程控制图

结合广东省实际的创新环境情况，改进型 EWMA 控制图判断出的四个离群值是更符合实际的。(1)2008 年底，中国受到美国严重的次贷危机影响，整个大环境的社会经济状态是属于一个下降(安全状况降低)的情况，受其影响的创新创业环境状态落在了控制线之外，这种情况更符合当时的实际情况。而且这种状态连续持续了三期，多了两个邻近的离群值，包括 2009 年及 2010 年，这几个离群值的存在都具有一定的合理性，无论是受到宏观政策、外部环境还是其他一系列经济因素的影响，宏观上创新环境状态的情况在一段时间内应具有一定的持续性，至少是短期的持续性。(2)2002 年，虽然广东省创新能力稳居全国第三位，总体水平较高，但是影响创新能力各要素的方法是不平衡的，其中包括整个创新环境状态。首先，2002 年广东省研发经费支出中，政府投入部分约占 10%，该数据为全国最低，投入明显不足，政府对科技创新的重视程度远远不够。同时，广东省产学研合作还处于一个发展的态势，企业与研究机构、高等院校合作仍不太紧密，特别在某些领域的合作发展方面，学科分布与企业需求不太相符，这也造成了创新环境状态的变动。需要说明的是，落在控制线外的样本数据为离群值，这里的离群值并不一定是指创新环境状况不好的情形，也可能是创新环境状态极好，却远远偏离其他数据的情形，此时应根据实际情况进行分析。

四、研究结论及启示

基于广东省高新技术企业 1997—2019 年的数据，本文从宏观角度研究创新环境状态的相关因素及指标体系、构建创新环境指数及进行预测预警分析，主要研究结论为：

第一，创新环境状态指标体系涵盖了经济环境及硬件能力、制度及创新人才因素、产业聚集等几个方面，构建出的创新环境指数综合反映出地区在这几个方面的创新环境状态。在对广东省创新环境各指标的实际数据进行变异系数测度时，发现波动程度较大（变异系数值＞2.0）的指标包括：两个经济因素指标（专业机构服务数增长率和入园企业数量增长率）、一个制度因素指标（高新区相关政策措施增长率）和三个产业聚集指标（高新区 GDP 增长率、高新技术产品产值增长率和产值超亿元企业数量增长率）。这几个指标的波动程度类似，也与它们之间的正向相互作用有关。随着高新区科技企业的投入及政策扶持力度加强，会吸引更多优势企业入园，相应带动专业机构服务数量增长；同时，优势企业入园又会提高资源共享及高新区的竞争力，从而带动整个高新区的 GDP 及技术产品产值增加。因此，为了提升整个高新技术企业的创新环境，政府可持续加大对高新区在各方面的支持力度，加大财政科技投入，继续实施相关税收优惠政策。同时，也应重视变异系数较小的指标，如本科以上学历占高新区从业人员比重，政府可通过制定人才奖励政策等，为高新区引进更多博士和硕士学历人员，加强创新人才队伍建设，这也有利于改善创新环境。

第二，构建了创新环境指数 MacroIEI，由于创新环境发展过程涉及的经济运行指标之间往往存在着相关关系，这里采用的是稳健主成分合成指数法。从图 2 的广东省创新环境指数走势图可以看出，MacroIEI 是波动的，且较大的波动后面往往跟着另一个较大的波动，较小的波动后面往往跟着另一个较小的波动，即：波动具有一定的持续性。从整体上看，样本考察期内的创新环境指数呈现出平缓上涨的趋势。

第三，利用自适应 ARIMA 方法对 MacroIEI 进行预测，并首次将预警监控技术引入到创新环境状态的研究分析中，利用改进型 EWMA 控制图对创新环境指数 MacroIEI 进行预警监控，与并未考虑样本数据相关性的常规 EWMA 控制图进行比较，发现常规控制图未识别出来的可能离群值，改进型 EWMA 控制图都识别出来了，且这几个离群值与实际情况较为相符。因此，相比常规控制图 EWMA 而言，改进型 EWMA 控制图具有更好的检测效果。

本研究得到的指数趋势及预警监控结论基本符合广东省实际的创新环境状态，在一定程度上验证了构建的创新环境指数的适用性，提出的动态预警监控方法也是合理的，可以依据这些来评价广东省的创新环境状态，进一步构建区域创新环境预警系统，本文采用的模型方法对中国其他地区的创新环境状态研究具有一定的借鉴意义。

参考文献：

[1] OECD. Innovative Networks：Cooperation in National Innovation Systems [R]. Paris：OECD，2001：7-41.

[2] 冯之浚. 国家创新系统的理论与政策[J]. 群言，1999(2)：22-29.

[3] 胡志坚，苏靖. 区域创新系统理论的提出与发展[J]. 中国科技论坛，1999(6)：20-23.

[4] 方匡南，陈子岚. 基于半监督广义可加 Logistic 回归的信用评分方法[J]. 系统工程理论与实践，2020，40(2)：392-402.

[5] 杨明海，张红霞，孙亚男，等. 中国八大综合经济区科技创新能力的区域差距及其影响因素研究[J]. 数量经济技术经济研究，2018(4)：3-19.

[6] 宋旭光，赵雨涵. 中国区域创新空间关联及其影响因素研究[J]. 数量经济技术经济研究，2018，35(7)：23-41.

[7] Porter M E，Furman J L ，Stern S. The Drivers of National Innovative Capacity：Implications for Spain and Latin America [J]. Harvard Business Review，2000(5)：1-13.

[8] Porter M E. On Competition [J]. Bottom Line，1998，13(1)：36-39.

[9] Martina Fromhold-Eisebith. Innovative Milieu and Social Capital Complementary or Redundant Concepts of Collaboration-based Regional Development [J]. European Planning Studies，2004，12(6)：747-765.

[10] Collins P，Pontikakis D. Innovation Systems in the European Periphery：The Case of Ireland and Greece [C]// European Regional Science Association，2006.

[11] 杨毅. 陕西创新指数研究[J]. 统计与信息论坛，2014，29(4)：70-74.

[12] 施雷，陈红亚. 陕西省创新创业环境评价及比较研究[J]. 中国科技资源导刊，2019(2)：28-34.

[13] 李玲，陶厚永. 包容性创新环境指标体系构建及区域比较[J]. 科技进步与对策，2016(24)：116-123.

[14] 崔航，曹扬，王维才. 我国典型城市群创新环境评价研究[J]. 中国管理信息化，2016(19)：136-140.

[15] 党晶晶，姚顺波，吴棋滢. 基于时空视角的区域 EES 系统协调发展评价研究——以榆林市为例[J]. 西北大学学报（哲学社会科学版），2014，44(3)：146-152.

[16] 黄玲雁. 基于因子分析的湖南省区域创新环境研究[J]. 南华大学学报(社会科学版)，2014，15(1)：59-63.
[17] 陈红川. 基于聚类分析的高新技术产业自主创新环境评价研究[J]. 广州大学学报(自然科学版)，2017(4)：83-88.
[18] 李沃源，乌兰. 基于系统构成要素的西部地区科技创新系统环境组合评价研究[J]. 科学管理研究，2019，37(1)：59-62.
[19] 于洋，朱建平，郭华生. 新战略背景下城市经济发展状况测度研究——基于融合社会网络的实证分析[J]. 统计研究，2021，38(3)：30-43.
[20] 苏炜，蔡丽茹. 珠三角城市创新环境评价及政策研究[J]. 科技管理研究，2018，38(10)：67-76.
[21] 林媛媛，张美月. 环境规制与技术创新协调发展的对策——基于福建省的分析[J]. 现代经济信息，2017(22)：460-461.
[22] 王斌会. 稳健主成分分析方法研究及其在经济管理中的应用[J]. 统计研究，2007(8)：74-78.
[23] 国家统计局社科文司"中国创新指数(CII)研究"课题组. 中国创新指数研究[J]. 统计研究，2014，31(11)：24-28.
[24] Hunter J S. The Exponentially Weighted Moving Average [J]. Journal of Quality Technology，1986，18(4)：19-25.
[25] Zhang N F. A Statistical Control Chart for Stationary Process Data [J]. Technometrics，1998，40(1)：24-38.

Innovation Environment Indicator Construction and Dynamic Early Warning Monitoring Research Based on Adaptive ARIMA Method

XIE Xian-fen[1a]，WANG Bin-hui[1b]，GU Wan-rong[2]，YANG Ying[3]

(a. School of Economics，b. School of Management，1. Jinan University，Guangzhou 510632，China；
2. School of Mathematics and Information，South China Agricultural University，Guangzhou 510642，China；
3. Law School，Central South University，Changsha 410083，China)

Abstract：Regional economic development has always attached great importance to the issue of innovative environment，and also one of the key contents of long-term research on regional innovation capability for scholars. Based on the reality of high-tech enterprises，combined with recent research results on relevant issues at home and abroad，an indicator system for the state of innovation environment has been established from five levels：economic factors，institutional factors，talent factors，hardware capabilities，and industrial agglomeration. Based on the annual data of high-tech enterprises in Guangdong from 1997 to 2019，constructing the innovation environment indicator MacroIEI with robust principal component composite index method. The analysis results show that the indicator value obtained is in good agreement with the actual situation. Then，using the adaptive ARIMA method to build the prediction model for innovation environment indicator，and adding the prediction results to improved EWMA control chart to realize the dynamic process control of innovation environment indicator. In order to verify the reliability of the model，It is compared with the conventional EWMA control chart. The analysis shows that in the early warning monitoring of innovation environment indicator MacroIEI，the improved EWMA control chart has better detection effect than the conventional EWMA control chart. Through the analysis of process control results of MacroIEI，the dynamic early warning and monitoring method proposed here is reasonable. On the basis，government decision-making departments or related enterprises can evaluate the state of regional innovation environment and make decisions.

Key words：innovation environment indicator；adaptive ARIMA；early warning monitoring；improved EWMA control chart

(责任编辑：李　勤)

计算机应用研究

- 第二届国家期刊奖百种重点期刊
- 中国期刊方阵双效期刊
- 中国科技核心期刊
- 全国中文核心期刊
- 中国计算机学会会刊
- 2020—2022年科技期刊世界影响力指数（WJCI）报告收录期刊
- 中国科学引文数据库（CSCD）来源期刊
- RCCSE中国核心学术期刊
- 中国科技期刊精品数据库全文来源期刊
- 中国学术期刊综合评价数据库来源期刊
- 《中国期刊网》《中国学术期刊（光盘版）》来源期刊
- 2017—2019年中国国际影响力优秀学术期刊（自然科学与工程技术）
- 中国精品科技期刊顶尖学术论文（F5000）项目来源期刊
- 《中国工程技术电子信息网》《电子科技文献数据库》来源期刊
- 英国《科学文摘》（INSPEC）来源期刊
- 《日本科学技术振兴机构数据库》（JST）来源期刊
- 俄罗斯《文摘杂志》（AJ, VINITI）来源期刊
- 美国《艾博思科学术数据库》（EBSCO）全文来源期刊
- 美国《剑桥科学文摘（自然科学）》（CSA（NS））核心期刊
- 波兰《哥白尼索引》（IC）来源期刊
- 美国《乌利希期刊指南（网络版）》（Ulrichsweb）收录期刊

ISSN 1001-3695
CN 51-1196/TP

10 >
9 771001 369236

计算机应用研究 第四十卷第十期 二〇二三年十月

ISSN 1001-3695
CODEN JYYIC7

计算机应用研究

Application Research of Computers

第40卷 第10期 2023年10月
Vol.40 No.10 Oct. 2023

10

2023

四川省计算机研究院主办
中国计算机学会会刊

❖英国《科学文摘》(INSPEC)来源期刊 ❖《日本科学技术振兴机构数据库》(JST)来源期刊
❖俄罗斯《文摘杂志》(AJ)来源期刊 ❖美国《艾博思科学术数据库》(EBSCO)全文来源期刊
❖美国《乌利希期刊指南(网络版)》(Ulrichsweb)收录期刊 ❖美国《剑桥科学文摘(自然科学)》(CSA(NS))核心期刊
❖2017—2019年中国国际影响力优秀学术期刊(自然科学与工程技术) ❖波兰《哥白尼索引》(IC)来源期刊
❖第二届国家期刊奖百种重点科技期刊 ❖中国科学引文数据库(CSCD)来源期刊
❖中国科技核心期刊 ❖全国中文核心期刊 ❖《中文科技期刊数据库》来源期刊
❖中国科技论文统计源期刊 ❖《中国期刊网》《中国学术期刊(光盘版)》来源期刊
❖中国学术期刊综合评价数据库来源期刊 ❖中国精品科技期刊顶尖学术论文(F5000)项目来源期刊
❖RCCSE核心学术期刊 ❖中国期刊方阵双效期刊 ❖《电子科技文献数据库》来源期刊
❖《中国工程技术电子信息网》来源期刊

计算机应用研究

Jisuanji Yingyong Yanjiu

第 40 卷 第 10 期 2023 年 10 月

目 次

综述评论

面向深度网络的小样本学习综述 …… 潘雪玲,李国和,郑艺峰(2881)

基于多种云环境的任务调度算法综述 …… 陈红华,崔翛龙,王耀杰(2889)

从扩展性角度看区块链 …… 王 锋,张 强,刘 扬,刘琳琳,路 阳(2896)

区块链技术

一种基于区块链的车联网安全认证协议 …… 包 俊,张新有,冯 力,邢焕来(2908)

基于节点分组信誉模型的改进 PBFT 共识算法 …… 陈苏明,王 冰,陈玉全,邢 涛,马宇辉,赵建立(2916)

基于超级账本的蚁群因子差分进化算法的可信服务组合优化 …… 冉瑞生,刘 震,祁 翔,彭顺顺(2922)

算法研究探讨

用于连续时间中策略梯度算法的动作稳定更新算法 …… 宋江帆,李金龙(2928)

基于自适应交互学习的 CPS 时间序列异常检测 …… 伍冠潮,凌 捷(2933)

基于相对熵和二元熵的多粒度直觉语言 TODIM 决策方法 …… 郭奉佳,高建伟,陈 炜(2939)

基于双层记忆网络的多领域端到端任务型对话系统 …… 方明弘,万 里,戴凡杰(2945)

融合一致性社交关系的协同相似嵌入推荐模型 …… 吴永庆,孙 鹏,金 尧,丁治辰(2951)

基于 Spark 和 AMPSO 的并行深度卷积神经网络优化算法 …… 刘卫明,罗全成,毛伊敏,彭 喆(2957)

基于图卷积网络融合依存信息的事件检测方法 …… 张紫月,王 羽,徐 建(2967)

基于时空残差张量学习的城市路网交通数据修复 …… 李金龙,李若南,吴 攀,于广婧,许伦辉(2972)

混合哈里斯鹰优化算法求解带模糊需求的低碳多式联运路径规划问题 …… 黄 琴,张惠珍,马 良,杨健豪(2978)

基于 Prim-DMGA 算法的闭环供应链网络鲁棒优化研究 …… 孙军艳,陈泽飞,陈智瑞,李晓朋(2984)

基于改进聚类算法的交通事故多发点识别方法 …… 王艺霖,肖媛媛,左鹏飞,杨 博,刘悦霞,段宗涛(2993)

引入动态分化和邻域诱导机制的双蚁群优化算法 …… 禹博文,游晓明,刘 升(3000)

考虑竞争环境的联邦学习多维契约激励机制研究 …… 杨 扬,殷红建,王 超(3007)

基于多重分形及其关联特征组合的心电信号分类 …… 卢 清,李秋生,叶莉华,许德鹏(3016)

面向图数据的结构化正则路径查询方法 …… 邱 涛,王屿涵,邓国鹏,孙 尧,吕光华,夏秀峰(3022)

基于产业互联网络节点创新要素的链路预测方法 …… 何建佳,廖耀文,周　洋(3028)

融合多策略的增强麻雀搜索算法及其应用 …… 李大海,李　鑫,王振东(3032)

融合评分上下文和物品相似度的推荐算法 …… 卢泽伦,古万荣,毛宜军,陈梓明(3040)

逐层 Transformer 在类别不均衡数据的应用 …… 杨晶东,李熠伟,江　彪,姜　泉,韩　曼,宋梦歌(3047)

基于熵权的全局记忆 LF 蚁群聚类算法 …… 熊伟超,蒋　瑜(3053)

系统应用开发

基于 GMEC 转换算法的 Petri 网结构控制器综合方法 …… 郁　希,黎　良(3059)

一种 TCN 的改进模型及其在短期光伏功率区间预测的应用 …… 宋绍剑,姜屹远,刘　斌(3064)

基于 PVC-CAE 的轴承剩余寿命预测方法 …… 张远亮,李海浪(3070)

软件技术研究

基于图注意力的异构图社交推荐网络 …… 吴相帅,孙福振,张文龙,张志伟,王绍卿(3076)

基于层次注意力机制的源代码迁移模型 …… 李　征,徐明瑞,吴永豪,刘　勇,陈　翔,武淑美,刘恒源(3082)

结合特征对齐与实例迁移的跨项目缺陷预测 …… 李　莉,赵　鑫,石可欣,苏仁嘉,任振康(3091)

基于一致性图卷积模型的多模态对话情绪识别 …… 谭晓聪,郭军军,线岩团,相　艳(3100)

网络与通信技术

双无线供能 IRS 辅助的无线通信系统的吞吐量优化研究 …… 陈健锋,崔　苗,张广驰(3107)

高效快速融合的双 LAN 太赫兹无线局域网 MAC 协议 …… 古金东,任　智,陈春宇,刘　洋(3114)

基于纠错反馈的 FCrSK 带内频分多址干扰对消算法 …… 康远鹏,杨　超,聂小红,郑　霖(3120)

基于任务需求匹配的网络连通质量控制方法 …… 何杏宇,王　怡,杨桂松,贾明权,周　玥,徐心玥(3125)

信息安全技术

基于 Tamarin 的 MQTT 协议安全性分析方法 …… 郑红兵,王焕伟,赵　琪,董姝岐,井　靖(3132)

基于流量切片的 DNS 隐蔽通道检测 …… 刘阳洋,阮树骅,曾雪梅(3138)

基于动态间隔压缩的鲁棒网络流水印算法 …… 马香港,李腾耀,刘粉林,冯王昕,杨春芳(3144)

基于 KFCMSA 的(k,l)加权社交网络匿名算法 …… 史　伟,王园园,李　刚,张　兴(3149)

图形图像技术

基于透视网格的自适应窄带表面粒子提取方法 …… 周志强,吴　桐,张严辞(3155)

基于对比记忆网络的弱监督视频异常检测 …… 李文中,吴克伟,孙永宣,焦　畅,熊思璇(3162)

基于先验知识的单视图三维点云重建算法研究 …… 陈雅丽,李海生,王晓川,李　楠(3168)

场景关系图学习的群组行为识别 …… 焦　畅,吴克伟,于　磊,谢　昭,李文中(3173)

基于多维注意力融合的驾驶场景分割增强算法 …… 刘奕晨,章坚武,胡　晶(3180)

物体显著性排名感知网络用于高效图像检索 …… 李林峰,陈程立诏,王恒森(3186)

融合比例先验和损失感知的目标检测模型的正负样本分配算法 …… 庄旭君,左华红,韩　屏(3194)

信息集萃

下期要目 …… (3179)

期刊基本参数:CN51-1196/TP * 1984 * m * A4 * 320 * zh * P * ￥50.00 * 3840 * 48 * 2023-10 * n

本期责任编辑　黄　莉,何　俐

Application Research of Computers
Vol. 40 No. 10 Oct. 2023

Contents

Survey

Survey on few-shot learning for deep network ····· Pan Xueling, Li Guohe, Zheng Yifeng(2881)

Summary of task scheduling algorithms based on multiple cloud environments ····· Chen Honghua, Cui Xiaolong, Wang Yaojie(2889)

Research progress of blockchain from perspective of scalability ····· Wang Feng, Zhang Qiang, Liu Yang, *et al*. (2896)

Technology of Blockchain

Security authentication protocol for Internet of Vehicles based on blockchain ····· Bao Jun, Zhang Xinyou, Feng Li, *et al*. (2908)

Improved PBFT consensus algorithm based on node grouping reputation model ····· Chen Suming, Wang Bing, Chen Yuquan, *et al*. (2916)

Ant colony factor differential evolutionary algorithm based on Hyperledger Fabric for trustworthy service composition optimization ····· Ran Ruisheng, Liu Zhen, Qi Xiang, *et al*. (2922)

Algorithm Research & Explore

Action stable updating algorithm for policy gradient methods in continuous time ····· Song Jiangfan, Li Jinlong(2928)

Time series anomaly detection for cyber physical systems based on adaptive interactive learning ····· Wu Guanchao, Ling Jie(2933)

Multi-granular intuitionistic linguistic TODIM method based on relative entropy and binary entropy ····· Guo Fengjia, Gao Jianwei, Chen Wei(2939)

Multi-domain end-to-end task-oriented dialogue system based on double-layer memory network ····· Fang Minghong, Wan Li, Dai Fanjie(2945)

Collaborative similarity embedding recommendation model incorporating consistent social relationships ····· Wu Yongqing, Sun Peng, Jin Yao, *et al*. (2951)

Parallel deep convolution neural network optimization algorithm based on Spark and AMPSO ····· Liu Weiming, Luo Quancheng, Mao Yimin, *et al*. (2957)

Event detection based on dependency information and graph convolutional network ····· Zhang Ziyue, Wang Yu, Xu Jian(2967)

Traffic data imputation of urban road network based on spatial-temporal residual tensor learning ····· Li Jinlong, Li Ruonan, Wu Pan, *et al*. (2972)

Hybrid Harris hawks optimization algorithm for solving low-carbon multimodal transportation problem with fuzzy demand ····· Huang Qin, Zhang Huizhen, Ma Liang, *et al*. (2978)

Robust optimization of closed-loop supply chain network based on Prim-DMGA algorithm ····· Sun Junyan, Chen Zefei, Chen Zhirui, *et al*. (2984)

Identifying method of traffic accident-prone spots based on improved clustering algorithm ····· Wang Yilin, Xiao Yuanyuan, Zuo Pengfei, *et al*. (2993)

Dual-ant colony optimization algorithm with dynamic differentiation and neighborhood induction mechanism ··· Yu Bowen, You Xiaoming, Liu Sheng(3000)

Incentive mechanism for federated learning in competitive environment: multidimensional contract approach ····· Yang Yang, Yin Hongjian, Wang Chao(3007)

ECG signal classification based on combination of multifractal and its correlation features ····· Lu Qing, Li Qiusheng, Ye Lihua, *et al*. (3016)

Efficient method of structured regular path query for graph data ····· Qiu Tao, Wang Yuhan, Deng Guopeng, *et al*. (3022)

Link prediction method based on nodal innovation elements of industrial interconnection network ····· He Jianjia, Liao Yaowen, Zhou Yang(3028)

Enhanced sparrow search algorithm with multiple strategies and its application ····· Li Dahai, Li Xin, Wang Zhendong(3032)

Recommendation algorithm with rating context and item similarity ····· Lu Zelun, Gu Wanrong, Mao Yijun, *et al*. (3040)

Application of layer by layer Transformer in class-imbalanced data ····· Yang Jingdong, Li Yiwei, Jiang Biao, *et al*. (3047)

Global memory LF ant colony clustering algorithm based on entropy weight ····· Xiong Weichao, Jiang Yu (3053)

System Development & Application

Synthesis of Petri net structural controller based on GMEC transformation algorithm ····· Yu Xi, Li Liang (3059)

Improved TCN model and its application in short-term photovoltaic power interval prediction ····· Song Shaojian, Jiang Yiyuan, Liu Bin (3064)

Remaining life prediction method of bearing based on PVC-CAE ····· Zhang Yuanliang, Li Hailang (3070)

Software Technology Research

GAT based heterogeneous graph neural network for social recommendation ····· Wu Xiangshuai, Sun Fuzhen, Zhang Wenlong, *et al*. (3076)

Source code migration model based on hierarchical attention mechanism ····· Li Zheng, Xu Mingrui, Wu Yonghao, *et al*. (3082)

Cross-project defect prediction combining feature alignment and instance migration ····· Li Li, Zhao Xin, Shi Kexin, *et al*. (3091)

Consistency based graph convolution network for multimodal emotion recognition in conversation ····· Tan Xiaocong, Guo Junjun, Xian Yantuan, *et al*. (3100)

Technology of Network & Communication

Throughput optimization for wireless communication systems assisted by dual wireless powered IRS ····· Chen Jianfeng, Cui Miao, Zhang Guangchi (3107)

Efficient and fast convergence of dual-LAN terahertz wireless LAN MAC protocol ····· Gu Jindong, Ren Zhi, Chen Chunyu, *et al*. (3114)

Interference cancellation algorithm of FCrSK in-band FDMA based on correction feedback ····· Kang Yuanpeng, Yang Chao, Nie Xiaohong, *et al*. (3120)

Network connectivity quality control method based on matching task requirement ····· He Xingyu, Wang Yi, Yang Guisong, *et al*. (3125)

Technology of Information Security

Tamarin-based security analysis of MQTT protocol ····· Zheng Hongbing, Wang Huanwei, Zhao Qi, *et al*. (3132)

DNS covert channel detection based on traffic slice ····· Liu Yangyang, Ruan Shuhua, Zeng Xuemei (3138)

Robust network flow watermarking algorithm based on dynamic interval compression ····· Ma Xianggang, Li Tengyao, Liu Fenlin, *et al*. (3144)

(k,l) weighted anonymous algorithm for social network based on KFCMSA ····· Shi Wei, Wang Yuanyuan, Li Gang, *et al*. (3149)

Technology of Graphic & Image

Perspective-grid-based adaptive narrow-band surface particle extraction method ····· Zhou Zhiqiang, Wu Tong, Zhang Yanci (3155)

Video anomaly detection combining with contrastive memory network ····· Li Wenzhong, Wu Kewei, Sun Yongxuan, *et al*. (3162)

Single-view 3D point cloud reconstruction algorithm based on priori knowledge ····· Chen Yali, Li Haisheng, Wang Xiaochuan, *et al*. (3168)

Scene relation graph network for group activity recognition ····· Jiao Chang, Wu Kewei, Yu Lei, *et al*. (3173)

Driving scene segmentation enhancement algorithm based on multidimensional attention fusion ····· Liu Yichen, Zhang Jianwu, Hu Jing (3180)

Object saliency ranking awareness network for efficient image retrieval ····· Li Linfeng, Chen Chenglizhao, Wang Hengsen (3186)

Positive and negative samples allocation algorithm for object detection models incorporating ratio-prior and loss-aware ····· Zhuang Xujun, Zuo Huahong, Han Ping (3194)

Responsible Editor Huang Li, He Li

融合评分上下文和物品相似度的推荐算法*

卢泽伦[1]，古万荣[1,2†]，毛宜军[1]，陈梓明[1]
（1. 华南农业大学 数学与信息学院，广州 510642；2. 广州市智慧农业重点实验室，广州 510642）

摘　要： 推荐系统中用户的评分往往会受到评分上下文的影响，即用户先前对一些物品的评分会影响其对当前物品评分的客观性。稀疏线性方法在计算物品相似度时将受到上下文影响的用户评分与其他评分同等看待，然而该部分评分并不能客观地反映出物品之间的相似度。针对以上问题，在稀疏线性方法的基础上提出了融合评分上下文和物品相似度的推荐算法，算法分为三个阶段：第一个阶段使用加权评分计算物品最近邻进行特征选择；第二个阶段利用评分误差权重减少算法模型对受到上下文影响的评分的拟合，训练得出物品相似度矩阵；第三个阶段根据用户评分和物品相似度进行评分预测以完成物品推荐。在 MovieLens 的四个数据集上进行实验，采用平均准确率（MAP）、平均倒数排名（MRR）和归一化折损累计增益（NDCG）指标来评估算法效果。实验结果表明，融合评分上下文将进一步提高物品相似度的准确性，从而提高推荐的性能。

关键词： 显式反馈；推荐系统；评分上下文；物品相似度；稀疏线性方法
中图分类号： TP391　　**文献标志码：** A　　**文章编号：** 1001-3695(2023)10-024-3040-07
doi:10.19734/j.issn.1001-3695.2023.02.0057

Recommendation algorithm with rating context and item similarity

Lu Zelun[1], Gu Wanrong[1,2†], Mao Yijun[1], Chen Ziming[1]
(1. *College of Mathematics & Informatics, South China Agricultural University, Guangzhou* 510642, *China*; 2. *Guangzhou Key Laboratory of Intelligent Agriculture, Guangzhou* 510642, *China*)

Abstract: In the recommendation system, the user's ratings are often affected by the rating context, that is, the user's previous ratings of some items will affect the objectivity of his rating of the current item. Sparse linear method treats user ratings affected by context as the same as other ratings when calculating item similarity. However, this partial ratings cannot objectively reflect the similarity between items. To solve the above problems, this paper proposed a recommendation algorithm combining rating context and item similarity based on sparse linear method. It divided the algorithm into three stages. The first stage used weighted ratings to calculate the item's nearest neighbor for feature selection. In the second stage, it used the rating error weight to reduce the fitting of the ratings affected by the context of the algorithm model, and trained the item similarity matrix. In the third stage, it predicted the ratings according to the user's ratings and the item similarity, and finally sorted the predicted ratings to complete the item recommendation. Experiments were conducted on four datasets of MovieLens, it used mean average precision (MAP), mean reciprocal rank (MRR) and normalized discounted cumulative gain (NDCG) to evaluate the effectiveness of the algorithm. The experimental results show that the fusion rating context will further improve the accuracy of item similarity and thus improve the performance of recommendation.

Key words: explicit feedback; recommendation system; rating context; item similarity; sparse linear method

0　引言

随着互联网的高速发展和数据信息的爆炸式增长，人们面临着严重的信息过载问题，推荐系统作为解决信息过载问题的主要技术之一，可根据用户的历史数据将物品准确地推荐给用户，同时获取用户对物品的评分，进一步分析出用户的偏好以进行更好的推荐[1~3]。用户的历史数据常分为隐式数据和显式数据，隐式数据包括用户对物品的点击等[4,5]，显式数据包括用户对物品的评分等[6,7]。然而在显式数据中，用户对物品的评分往往不完全代表用户的实际偏好或者物品的质量，该评分还受到其他因素的影响[8]。在很多情况下，用户先前对其他物品的评分情况会影响其对当前物品的评分，如在观看电影的场景中，当用户先前观看的电影的质量很低且用户对该电影的评分也很低，那么用户很有可能会给当前观看的电影评一个高分，即使当前电影的质量一般，这种情况称为用户的评分受到评分上下文的影响。例如，用户 A 对剧情电影 A 的评分原本只能打到 3 分，然而因为用户 A 在观看剧情电影 A 之前看了一部十分差劲的剧情电影 B，导致其在两部电影间作出对比之下给剧情电影 A 打了 5 分的高分。在包括稀疏线性方法（sparse linear method，fsSLIM）[9]等使用物品相似度进行推荐的推荐算法中，计算物品相似度常基于用户对物品的评分，且将用户的所有评分同等对待[10~12]，忽略了受到上下文影响的用户主观评分并不能客观地反映出物品之间的相似度这一问题，即物品之间的相似度不应受到用户评分先后顺序的影响。针对上述问题，本文在稀疏线性方法的基础上提出一种利用时间信息来融合评分上下文和物品相似度的推荐算法（fsSLIM with rating context，RCfsSLIM）。首先，利用时间信息对用户评

收稿日期： 2023-02-18；**修回日期：** 2023-04-20　**基金项目：** 中山大学广东省计算科学重点实验室开放基金资助项目（2021010）；广东省自然科学基金面上项目（2022A1515011489）；国家社科基金后期资助项目（19FTJB001）；广东省哲学社会科学规划项目（GD19CGL34）

作者简介： 卢泽伦（1997-），男，广东惠州人，硕士，CCF 会员，主要研究方向为推荐系统、机器学习；古万荣（1982-），男（通信作者），广东梅州人，讲师，硕导，博士，主要研究方向为搜索引擎、互联网大数据分析与挖掘、推荐系统（guwanrong@scau.edu.cn）；毛宜军（1979-），男，湖北公安人，讲师，博士，主要研究方向为基因数据处理、信息安全、人工智能；陈梓明（1998-），男，广东广州人，硕士，主要研究方向为机器学习.

分进行加权处理,使用加权处理后的评分计算并选择出物品最近邻;然后,在基于物品最近邻的基础上,使用评分误差权重减少算法模型对受到上下文影响的评分的拟合,训练得出物品相似度矩阵,再通过结合用户评分矩阵和物品相似度矩阵计算得到预测评分,最终完成物品的推荐;最后,本文在四个标准数据集上进行了实验比较,结果表明融合评分上下文将进一步提高物品相似度的准确性,从而提高推荐的性能。

1　相关工作

本文提出的基于物品相似度的推荐算法利用时间信息将评分上下文融入其中,与之相关的工作包括用户评分上下文的影响、融合时间信息的推荐算法、基于物品相似度的推荐算法三方面。

1.1　用户评分上下文的影响

Lu 等人[8]统计用户阅读新闻的数据发现,用户的评分经常发生变化,且这些变化与用户交互的上下文高度相关。其将数据分为阅读前、阅读后和任务后。前两个阶段是在用户浏览有顺序的新闻列表上下文中收集的,这可能会受到新闻在列表中的位置、周围新闻和以前阅读的新闻的影响。为了消除这些影响,任务后阶段在用户完成新闻阅读后,将所有新闻重新显示给用户,然后再一次要求用户提供对每个新闻的实际评分,文献[8]认为此阶段中新闻在列表中的位置和上下文的影响都被消除了,并以此当做用户对每个新闻的实际评分。将收集的数据经过 T 检验后分析发现,用户对前一个点击的新闻的评分会影响当前阅读新闻的用户体验,当用户先前阅读不喜欢的新闻时,其在阅读后阶段中对当前阅读的新闻的评分会高于实际评分。分析数据表明,用户交互上下文将影响用户对当前阅读的新闻的即时评分,这可能与用户的实际评分不一致。因此本文利用时间信息体现出评分的上下文关系,将同一用户的评分按时间顺序排序后,对受到评分上下文影响的评分进行加权处理以改变其对算法模型的影响。

1.2　融合时间信息的推荐算法

传统的推荐系统倾向于利用用户对物品的所有评分来学习用户的长期偏好,在这种情况下用户的全部评分是同等重要的。然而用户对物品的选择往往受其最近的短期偏好以及最近浏览或购买物品等时间相关的上下文场景所影响[11,13]。针对上述问题,研究人员将时间信息融合进推荐算法中[14],近年来,基于序列和基于会话的推荐系统悄然兴起,其中一项重要的工作是处理物品中如时间、用户等上下文信息。Tang 等人[15]认为用户最近的偏好对其当前如何选择物品会造成更大的影响,于是通过卷积神经网络对用户短期序列的信息进行提取。Kang 等人[16]和 Sun 等人[17]使用自注意力机制对用户最近的若干个历史行为进行信息提取,以此将时间信息融合进推荐算法中。这些算法都表明融合了时间信息的推荐模型能有效地获得更好的推荐结果,也从另一角度表明融合时间信息有助于模型构建用户偏好或物品属性。因此,本文利用时间信息将评分上下文融合到推荐算法中,以获得更好的推荐效果。

1.3　基于物品相似度的推荐算法

推荐系统主要分为基于内容的推荐、基于协同过滤的推荐和混合推荐[18,19]三大类。结合用户评分和物品相似度进行推荐是基于协同过滤的推荐中常用的一种技术[11,20]。Kabbur 等人[21]通过用户评分训练出两个低维的隐因子矩阵,同时将两个隐因子矩阵乘积得到的高维矩阵视为物品的相似度矩阵。Lin 等人[22]将自注意力机制融合到算法中以获得更加准确的物品相似度。Ning 等人[9]提出了 fsSLIM 模型,该模型可通过基于邻域的方法提前计算出物品的相似物品进行特征选择。对于某一物品,模型只使用其相似物品对其进行学习,而不需要考虑其他所有物品,其目的是为了减小模型的训练时间。在特征选择步骤中,常用皮尔森系数度量相似度,物品 u 和 v 的皮尔森相似度计算方法如下:

$$\mathrm{sim}(u,v)=\frac{\sum_{i\in I_{uv}}(R_{i,u}-\overline{R_u})(R_{i,v}-\overline{R_v})}{\sqrt{\sum_{i\in I_{uv}}(R_{i,u}-\overline{R_u})^2}\sqrt{\sum_{i\in I_{uv}}(R_{i,v}-\overline{R_v})^2}} \tag{1}$$

其中:$R_{i,u}$和 $R_{i,v}$分别表示用户 i 对物品 u 的评分和用户 i 对物品 v 的评分;$\overline{R_u}$和$\overline{R_v}$分别表示物品 u 所有评分的平均分和物品 v 所有评分的平均分;I_{uv}表示对物品 u 和 v 都有评分的用户。接着 fsSLIM 模型可使用评分矩阵结合特征选择训练出物品的相似度矩阵,其目标函数如下:

$$\underset{W_j}{\text{minimize}}\frac{1}{2}\|R_j-\boldsymbol{R}W_j\|_2^2+\frac{\beta}{2}\|W_j\|_2^2+\lambda\|W_j\|_1 \tag{2}$$
$$W_j\geqslant 0,W_{j,j}=0$$

其中:R_j 是评分矩阵 $\boldsymbol{R}$ 的第 j 列;W_j 是物品相似度矩阵 $\boldsymbol{W}$ 的第 j 列。最后 fsSLIM 模型使用用户评分和物品相似度进行评分预测以进行物品推荐。然而,无论是在特征选择阶段还是模型训练阶段,fsSLIM 模型都忽略了用户的评分上下文对部分评分的影响,降低了训练得到的物品相似度矩阵的准确性和推荐效果。因此,本文在 fsSLIM 模型的基础上利用时间信息对用户评分进行加权处理,将受到上下文影响的评分与其他评分进行区分,并在特征选择阶段使用加权评分更加准确地计算出物品最近邻。同时在模型训练阶段根据评分的权重情况减少算法模型对受到上下文影响的评分的拟合,以此将评分上下文融入其中,从而提高推荐的性能。

2　融合评分上下文和物品相似度的推荐算法

本文所提出的 RCfsSLIM 算法框架如图 1 所示,主要分为三个阶段:在第一个阶段即特征选择阶段中,使用加权评分计算物品最近邻进行特征选择以减少算法模型训练时间;在第二个阶段即模型训练阶段中,使用评分误差权重减少算法模型对受到上下文影响的评分的拟合,以减少该评分对物品相似度的影响,训练得出物品相似度矩阵;在第三个阶段即评分预测与物品推荐阶段中,通过用户评分矩阵和物品相似度矩阵完成评分预测,最终将预测的评分进行排序后形成推荐的物品列表。

2.1　符号定义

在提出本文模型之前,先对本文主要使用到的三个矩阵进行说明:

矩阵 $\boldsymbol{R}$ 为用户—物品评分矩阵,即用户的评分矩阵。评分矩阵的每一行代表一个用户,每一列代表一个物品。评分矩阵第 u 行第 i 列的值 R_{ui}表示用户 u 对物品 i 的评分。

矩阵 $\boldsymbol{A}$ 是评分权重矩阵,大小与评分矩阵 $\boldsymbol{R}$ 相同。在评分权重矩阵 $\boldsymbol{A}$ 中,每一个位置上不为 0 的值 A_{ui}是对应评分矩阵 $\boldsymbol{R}$ 中评分 R_{ui}的权值,权值大小经过以下处理:a)对同一用户的评分按时间先后进行排序,若前一个评分等于最高评分,则当前评分权值大于 1;b)对同一用户的评分按时间先后进行排序,若前一个评分等于最低评分,则当前评分权值小于 1;c)其他情况下评分权值等于 1。其中,加权评分不应高于原始评分的更高一级评分且不应低于原始评分的更低一级评分,如用户评分为 4 分(评分区间为 1 ~5 分,评分间隔为 1 分),则加权评分不超过 5 分且不低于 3 分。

矩阵 $\boldsymbol{W}$ 为物品相似度矩阵,其行数与列数都和物品数量

相等,通过学习用户—物品评分矩阵得出。矩阵 $\boldsymbol{W}$ 的每一行或是每一列都代表一种物品,且 $\boldsymbol{W}$ 的每一列可视为此物品和此列对应物品的相似度。相似度矩阵第 u 行第 i 列的值 W_{ui} 表示第 u 个和第 i 个物品的相似度。

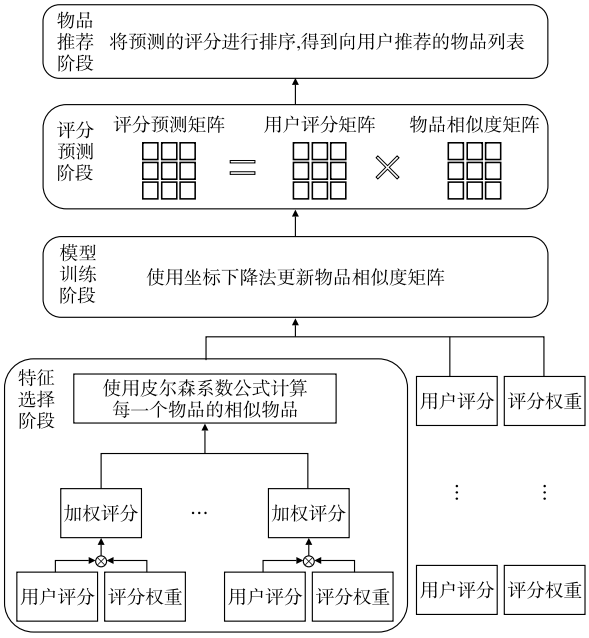


图 1 本文算法的整体框架

Fig. 1 Overall framework of proposed algorithm

2.2 RCfsSLIM 的模型

RCfsSLIM 模型计算用户 i 对未评分物品 j 的预测评分公式如下:

$$R_{ij}' = R_i W_j \tag{3}$$

其中:R_i 表示用户 i 对所有物品的评分;W_j 表示物品 j 与其他所有物品的相似度。例如,假设用户 i 对所有物品的评分为

$$R_i = (R_{i1}, R_{i2}, \cdots, R_{in}) \tag{4}$$

物品 j 与其他所有物品的相似度为

$$W_j = (W_{j1}, W_{j2}, \cdots, W_{jn})^{\mathrm{T}} \tag{5}$$

其中:n 为物品数量;R_{i1} 表示用户 i 对第一个物品的评分;W_{j1} 表示物品 j 与第一个物品的相似度,依此类推。若用户 i 对物品 j 的评分未知,则计算用户 i 对物品 j 的推荐评分为

$$R_{ij}' = R_{i1} \times W_{j1} + R_{i2} \times W_{j2} + \cdots + R_{in} \times W_{jn} \tag{6}$$

由上述可知,使用 RCfsSLIM 模型进行推荐的关键在于训练得到一个准确性高的物品相似度矩阵。

RCfsSLIM 模型可通过用户—物品的 $m \times n$ 评分矩阵 $\boldsymbol{R}$ 学习出一个 $n \times n$ 的物品相似度矩阵 $\boldsymbol{W}$。矩阵 $\boldsymbol{W}$ 的每一行或是每一列都代表一种物品,且 $\boldsymbol{W}$ 的每一列可视为此物品和此列对应物品的相似度。对于物品相似度矩阵 $\boldsymbol{W}$ 中每一列,RCfsSLIM 模型沿用 fsSLIM 模型的目标函数,并将其改写成:

$$\min_{W_j}\left[\sum_{i=1}^{m}(R_{i,j} - R_i W_j)^2 + \frac{\beta}{2}\|W_j\|_2^2 + \lambda\|W_j\|_1\right] \tag{7}$$

$$W_j \geqslant 0, W_{j,j} = 0$$

其中:m 是用户总数量;$R_{i,j}$ 表示用户 i 对物品 j 的评分;R_i 是评分矩阵 $\boldsymbol{R}$ 的第 i 行,表示用户 i 对所有物品的评分;W_j 是物品相似度矩阵 $\boldsymbol{W}$ 的第 j 列,表示物品 j 与其他所有物品的相似度;$\|\cdot\|_2$ 是向量的 L_2 范数,用于防止过拟合;$\|\cdot\|_1$ 是向量的 L_1 范数,保证向量的稀疏性;L_1 和 L_2 范数共同组成了弹性网络;常数 β 和 λ 是正则化参数;约束条件 $W_j \geqslant 0$ 表示该向量的每一个值都大于或等于 0,即考虑该列表示的物品与其他物品的相似度,而不考虑不相似度;约束条件 $W_{j,j} = 0$ 表示物品对自己的相似度为 0,若不限定此值为 0,则物品倾向于推荐自己,这样的结果无意义。该目标函数可以使用坐标下降法和软阈值函数来进行优化[23]。

2.2.1 使用加权评分进行特征选择

从 RCfsSLIM 模型的目标函数即式(7)可看出,对于评分矩阵 $\boldsymbol{R}$ 的第 j 列,该列是函数中的因变量,评分矩阵的其他 $n-1$ 列是函数中的自变量,即通过评分矩阵其余的 $n-1$ 列对第 j 列进行拟合。在这种情况下,可通过特征选择来减少评分矩阵中第 j 列的自变量列数,以减小训练时间。RCfsSLIM 模型提前计算出与评分矩阵第 j 列相似度高的其他 k 列,即与物品 j 最相似的 k 个物品,在最小化目标函数的时候只需使用该 k 列对第 j 列进行拟合。

特征选择中物品之间的相似度通过用户对物品的加权评分 R_{ui}^w 计算得出,即 $R_{ui}^w = R_{ui} * A_{ui}$。使用加权评分计算相似度的方法如下:

$$\mathrm{sim}(u,v) = \frac{\sum_{i \in I_{uv}}(R_{i,u}^w - \overline{R_u^w})(R_{i,v}^w - \overline{R_v^w})}{\sqrt{\sum_{i \in I_{uv}}(R_{i,u}^w - \overline{R_u^w})^2}\sqrt{\sum_{i \in I_{uv}}(R_{i,v}^w - \overline{R_v^w})^2}} \tag{8}$$

其中:$R_{i,u}^w$ 表示用户 i 对物品 u 的加权评分;$R_{i,v}^w$ 表示用户 i 对物品 v 的加权评分;$\overline{R_u^w}$ 表示物品 u 中所有加权评分的平均分;$\overline{R_v^w}$ 表示物品 v 中所有加权评分的平均分;I_{uv} 表示对物品 u 和 v 都有评分的用户。

2.2.2 使用坐标下降法优化 RCfsSLIM 目标函数

坐标下降法用于解决含有 L_1 范数或 L_2 范数或两者混合(弹性网络)的凸函数优化问题。弹性网络问题如下:

$$\min_{(\beta_0,\boldsymbol{\beta}) \in \mathbb{R}^{p+1}} R_\lambda(\beta_0, \boldsymbol{\beta}) =$$

$$\min_{(\beta_0,\boldsymbol{\beta}) \in \mathbb{R}^{p+1}}\left[\frac{1}{2N}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^{\mathrm{T}}\boldsymbol{\beta})^2 + \lambda P_\alpha(\boldsymbol{\beta})\right] \tag{9}$$

且

$$P_\alpha(\boldsymbol{\beta}) = (1-\alpha)\frac{1}{2}\|\boldsymbol{\beta}\|_{l_2}^2 + \alpha\|\boldsymbol{\beta}\|_{l_1} =$$

$$\sum_{j=1}^{p}\left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|\right] \tag{10}$$

其中:响应变量 $Y \in \mathbb{R}$;预测向量 $\boldsymbol{X} \in \mathbb{R}^p$;线性回归模型设为 $E(Y|\boldsymbol{X}=x) = \beta_0 + x\boldsymbol{\beta}$;$N$ 为观测对 (x_i, y_i) 的数量;p 是向量 $\boldsymbol{\beta}$ 的长度;P_α 是弹性网络,即 L_1 范数($\alpha=1$)和 L_2 范数($\alpha=0$)的组合。

求解式(9)可使用坐标下降法迭代更新每个 $\boldsymbol{\beta}$ 直到收敛,$\boldsymbol{\beta}$ 的更新形式如下:

$$\tilde{\beta}_j \leftarrow \frac{S\left(\frac{1}{N}\sum_{i=1}^{N}x_{ij}(y_i - \tilde{y}_i^{(j)}), \lambda\alpha\right)}{\frac{1}{N}\sum_{i=1}^{N}x_{ij}^2 + \lambda(1-\alpha)} \tag{11}$$

其中:$\tilde{y}_i^{(j)} = \tilde{\beta}_0 + \sum_{l \neq j}x_{il}\tilde{\beta}_l$;$S(z,y)$ 是软阈值函数。

$$S(z,y) = \mathrm{sign}(z)(|z| - y)_+ = \begin{cases} z - y & \text{if } z > 0 \text{ and } y < |z| \\ z + y & \text{if } z < 0 \text{ and } y < |z| \\ 0 & \text{if } y \geqslant |z| \end{cases} \tag{12}$$

结合式(9)和(10),可将 RCfsSLIM 模型的目标函数即式(7)转换为

$$\min_{W_j}\left[\frac{1}{2m}\sum_{i=1}^{m}(R_{i,j} - R_i W_j)^2 + \lambda\sum_{r=1}^{n}\left[\frac{1}{2}(1-\alpha)W_{r,j}^2 + \alpha|W_{r,j}|\right]\right] \tag{13}$$

$$W_j \geqslant 0, W_{j,j} = 0$$

结合式(11),坐标下降法更新物品相似度矩阵 $\boldsymbol{W}$ 的形式如下:

$$\tilde{W}_{r,j} = \frac{S\left(\frac{1}{m}\sum_{i=1}^{m}R_{ir}(R_{i,j} - \tilde{R}_{i,j}^{(r)}), \lambda\alpha\right)}{\frac{1}{m}\sum_{i=1}^{m}R_{ir}^2 + \lambda(1-\alpha)} \tag{14}$$

其中:$\tilde{R}_{i,j}^{(r)}=\sum_{l\neq r}R_{il}\tilde{W}_{l,j}$表示该次迭代中用户 i 对物品 j 的预测评分,则 $R_{i,j}-\tilde{R}_{i,j}^{(r)}$ 表示该次迭代中的评分误差;$S(z,y)$ 为软阈值函数,即式(12);在式(13)和(14)中的 $\lambda(1-\alpha)$ 和 $\lambda\alpha$ 分别代表式(7)中的 β 和 λ。

从式(14)可看出,在使用坐标下降法迭代更新物品相似度矩阵中的值时,评分误差起到了关键作用,即评分误差影响着物品相似度矩阵中的值。所以 RCfsSLIM 模型在计算用户每一个评分与预测评分的误差中,若当前评分的权重不为 1,则当前评分的评分误差权重 δ 小于 1,即 RCfsSLIM 推荐模型将减小对该评分的拟合以减小其对物品相似度矩阵 $\boldsymbol{W}$ 的影响,从而提高物品相似度矩阵 $\boldsymbol{W}$ 的客观性。将式(14)中迭代更新物品相似度矩阵时评分误差计算公式 $error_{ij}=R_{i,j}-\tilde{R}_{i,j}^{(r)}$ 改进为

$$error_{ij}'=\delta_{ij}(R_{i,j}-\tilde{R}_{i,j}^{(r)}) \tag{15}$$

其中:$R_{i,j}$表示用户评分;δ_{ij}表示该评分误差的权重;$\tilde{R}_{i,j}^{(r)}$ 表示该评分的预测评分,且 $\tilde{R}_{i,j}^{(r)}=\sum_{l\neq r}R_{il}\tilde{W}_{l,j}$。则 RCfsSLIM 模型使用坐标下降法更新物品相似度矩阵的最终形式如下:

$$\tilde{W}_{r,j}=\frac{S\left(\frac{1}{m}\sum_{i=1}^{m}R_{ir}\times\delta_{ij}(R_{i,j}-\tilde{R}_{i,j}^{(r)}),\lambda\alpha\right)}{\frac{1}{m}\sum_{i=1}^{m}R_{ir}^2+\lambda(1-\alpha)} \tag{16}$$

2.2.3　评分预测与物品推荐

迭代更新物品相似度矩阵 $\boldsymbol{W}$ 的值后,用户 i 对未评分物品 j 的预测评分计算方式如下:

$$R_{ij}'=R_iW_j \tag{17}$$

其中:$R_{ij}=0$;R_i 是评分矩阵第 i 行;W_j 是物品相似度矩阵第 j 列。因此 RCfsSLIM 计算出的评分矩阵如下:

$$\boldsymbol{R}'=\boldsymbol{RW} \tag{18}$$

排序由式(18)得到的预测评分,即可获得向用户推荐的物品列表。RCfsSLIM 模型的算法描述如下:

输入:用户的评分数据,每一评分数据包括用户 ID、物品 ID、评分值和时间戳。

输出:用户的物品推荐列表。

a)使用每一评分数据的时间戳信息计算出该评分的评分权重值。同时将每一评分数据的评分值和评分权重值相乘,得出加权评分。

b)根据式(8)使用加权评分计算出物品之间的相似度,得到物品的最近邻。

c)初始化物品相似度矩阵,根据式(16)使用物品的最近邻、评分值和评分权重,通过坐标下降法更新物品相似度矩阵。

d)物品相似度矩阵更新完成后利用式(18)计算得出评分预测矩阵。

e)对预测的评分进行排序,取预测评分高的多个物品作为物品推荐列表。

3　实验结果与分析

3.1　数据集

实验采用 MovieLens 的四个数据集,分别是 ml-100k、ml-latest-small(ml-ls)、Hetrec2011-movielens-2k(ml-10m2k)和 ml-1m。其中,Hetrec2011-movielens-2k 数据集将 MovieLens 数据集的电影与其在 IMDb 和烂番茄中的相应网页链接起来,是 MovieLens10M 数据集的扩展。数据集的具体信息如表 1 所示。

表 1　数据集具体信息

Tab. 1　Dataset specific information

数据集	用户数	物品数	评分数	评分间隔	稀疏度/%
ml-100k	943	1 682	100 000	1	93.70
ml-ls	610	9 742	100 836	0.5	98.30
ml-10m2k	2 113	10 197	855 598	0.5	96.03
ml-1m	6 040	3 900	1 000 209	1	95.75

3.2　评价指标

实验采用 top-N 推荐评价指标来评估推荐算法的性能,其中包括平均准确率(mean average precision,MAP)、平均倒数排名(mean reciprocal rank,MRR)和归一化折损累计增益(normalized discounted cumulative gain,NDCG)。

3.3　实验结果与分析

本文实验在四个数据集上都随机抽取了 70% 的训练数据和 30% 的测试数据。由于数据集 ml-ls 和 ml-10m2k 的评分间隔为 0.5 分,数据集 ml-100k 和 ml-1m 的评分间隔为 1 分,所以这四个数据集的评分权值进行如下设置:a)对同一用户的评分按时间先后进行排序,若前一个评分等于最高评分,则在数据集 ml-ls 和 ml-10m2k 中,当前评分的权值设为 1.1,在数据集 ml-100k 和 ml-1m 中,当前评分的权值设为 1.2;b)对同一用户的评分按时间先后进行排序,若前一个评分等于最低评分,则在数据集 ml-ls 和 ml-10m2k 中,当前评分的权值设为 0.9,在数据集 ml-100k 和 ml-1m 中,当前评分的权值设为 0.8;c)其他情况下评分权值等于 1。

除了 fsSLIM 模型以外,本文提出的 RCfsSLIM 模型还与以下五种基线模型进行了对比:

a) ItemKNN[10]:一种经典的基于物品的推荐算法,ItemKNN 根据物品之间的相似度大小进行排序后完成推荐。

b) WRMF[24,25]:该算法使用加权的方法来表征用户对物品行为的置信度,以此改变部分训练数据对模型的影响。

c) FISM[21]:该算法通过用户评分训练出两个低维的隐因子矩阵,两个矩阵乘积得到的高维矩阵视为物品的相似度矩阵。

d) LRML[26]:该算法使用用户和物品之间的潜在关系进行建模,以此提高特征选择的能力和推荐的性能。

e) AutoInt[27]:该算法使用多头自注意力来进行特征的交叉学习,有效地构造了高阶特征以获得更好的推荐效果。

3.4　实验结果与分析

3.4.1　推荐结果比较

本文提出的 RCfsSLIM 模型与六个基线模型应用于稀疏度不同的数据集上的整体效果如表 2 所示。本文算法 RCfsSLIM 明显优于其他六种对比方法,相比于 fsSLIM 模型,RCfsSLIM 模型在特征选择阶段使用加权评分计算物品相似度以找出更加准确的相似物品,并在模型训练时减小对部分受到用户评分上下文影响的评分的拟合,提高了训练得到的物品相似度矩阵的准确性,体现出更好的推荐效果。

3.4.2　第一个阶段中使用加权评分进行特征选择的作用

RCfsSLIM 模型第一个阶段中特征选择的本质是使用加权评分提前计算出物品之间的相似度,接着只使用物品的多个相似物品进行训练即可,由此可见在特征选择阶段中更加准确地找出某一物品的多个相似物品的重要性。

本节实验使用物品最近邻算法(item K-nearest neighbor,ItemKNN)探究使用加权评分进行特征选择的作用。ItemKNN 算法首先使用式(1)通过评分矩阵 $\boldsymbol{R}$ 中的物品向量来计算得到物品之间的相似度;接着对于一个用户未评分的物品,该算法将该物品和用户已评分物品的相似度进行累加;最后对每一个未评分物品累加的相似度进行排序从而达到推荐的目的。从相似度计算式(1)可以看出,计算物品相似度完全基于用户评分,即对于两个物品而言,若用户评分越相近,则两个物品的相似度越高,且 ItemKNN 在推荐物品的过程中仅累加物品之间的相似度进行排序,所以可将使用该算法进行推荐的性能视为物品之间相似度的准确性。

作为与式(1)的对比,本文在 ItemKNN 算法中使用式(8)通过加权评分计算物品之间的相似度,除此之外其余所有参数

保持不变。实验结果如表 3 所示,使用加权评分计算物品之间的相似度表现出更好的推荐效果,即对于某一个物品而言,使用加权评分可以更加准确地寻找出与其相似的 k 个物品,这也正是 RCfsSLIM 模型中特征选择阶段所需要的效果。

表 2 结果对比

Tab. 2 Comparison of results

数据集	top-N	评估指标	ItemKNN	WRMF	fsSLIM	FISM	LRML	AutoInt	RCfsSLIM
ml-100k	3	MAP	0.362 8	0.385 9	0.407 7	0.409 6	0.411 7	0.417 9	0.423 7
		MRR	0.596 4	0.631 1	0.648 1	0.649 5	0.651 5	0.660 9	0.666 7
		NDCG	0.440 9	0.463 1	0.482 9	0.487 6	0.488 4	0.491 2	0.499 3
	5	MAP	0.290 3	0.330 9	0.348 6	0.348 1	0.348 3	0.355 0	0.362 7
		MRR	0.597 6	0.649 8	0.667 1	0.672 4	0.673 5	0.678 1	0.685 8
		NDCG	0.393 9	0.439 4	0.454 9	0.455 3	0.459 6	0.462 8	0.469 7
ml-ls	3	MAP	0.283 7	0.329 5	0.349 2	0.350 2	0.352 6	0.354 5	0.360 8
		MRR	0.480 3	0.544 5	0.562 6	0.568 0	0.571 4	0.578 6	0.582 5
		NDCG	0.351 0	0.400 9	0.416 5	0.418 0	0.423 5	0.425 5	0.432 5
	5	MAP	0.243 1	0.266 1	0.287 3	0.290 2	0.295 0	0.295 9	0.299 7
		MRR	0.491 5	0.559 8	0.580 8	0.587 6	0.589 6	0.593 0	0.598 7
		NDCG	0.336 3	0.370 1	0.385 6	0.390 4	0.389 4	0.392 4	0.398 6
ml-10m2k	3	MAP	0.521 2	0.587 7	0.610 8	0.611 2	0.617 6	0.618 9	0.623 8
		MRR	0.670 1	0.744 4	0.773 2	0.778 5	0.779 2	0.783 6	0.788 0
		NDCG	0.589 5	0.641 5	0.665 9	0.665 8	0.669 7	0.672 8	0.678 2
	5	MAP	0.471 2	0.532 0	0.554 1	0.556 1	0.562 0	0.561 4	0.565 8
		MRR	0.682 5	0.755 2	0.781 0	0.779 2	0.782 9	0.788 5	0.794 9
		NDCG	0.548 6	0.612 4	0.631 3	0.636 0	0.639 2	0.641 4	0.647 2
ml-1m	3	MAP	0.276 6	0.339 8	0.365 0	0.366 1	0.373 0	0.375 5	0.380 3
		MRR	0.509 1	0.573 0	0.607 1	0.608 9	0.612 8	0.615 6	0.620 8
		NDCG	0.350 6	0.409 8	0.446 2	0.448 0	0.451 2	0.455 0	0.461 1
	5	MAP	0.213 0	0.276 3	0.301 3	0.304 6	0.308 5	0.309 7	0.315 4
		MRR	0.479 1	0.549 5	0.579 9	0.580 4	0.585 6	0.587 1	0.592 6
		NDCG	0.307 6	0.370 3	0.396 1	0.398 7	0.402 2	0.405 2	0.410 7

3.4.3 第二个阶段中评分误差权重的影响

RCfsSLIM 模型的第二个阶段中,一个评分的评分误差权重 δ 的大小影响着模型对该评分的拟合程度。若 δ 为 1,则表示模型对该评分进行完全拟合,仅在特征选择时使用加权评分计算物品相似度;若 δ 为 0,则表示模型不对该评分进行拟合;δ 的值在 0 ~ 1,则表示模型将减小对该评分的拟合。可见评分误差权重 δ 在 RCfsSLIM 推荐模型中的重要性。

表 3 ItemKNN 中加权评分与非加权评分结果对比

Tab. 3 Comparison of weighted and unweighted ratings results in ItemKNN

数据集	top-N	评估指标	非加权	加权	提升/%
ml-100k	3	MAP	0.362 8	0.385 6	6.28
		MRR	0.596 4	0.634 0	6.30
		NDCG	0.440 9	0.463 5	5.12
	5	MAP	0.290 3	0.303 4	4.52
		MRR	0.597 6	0.628 6	5.19
		NDCG	0.393 9	0.405 7	2.99
ml-ls	3	MAP	0.283 7	0.290 8	2.50
		MRR	0.480 3	0.501 0	4.30
		NDCG	0.351 0	0.364 0	3.70
	5	MAP	0.243 1	0.253 7	4.38
		MRR	0.491 5	0.511 6	4.09
		NDCG	0.336 3	0.348 5	3.63
ml-10m2k	3	MAP	0.521 2	0.536 4	2.91
		MRR	0.670 1	0.695 4	3.78
		NDCG	0.589 5	0.606 5	2.89
	5	MAP	0.471 2	0.483 9	2.70
		RR	0.682 5	0.717 7	5.16
		NDCG	0.548 6	0.570 1	3.91
ml-1m	3	MAP	0.276 6	0.285 4	3.19
		MRR	0.509 1	0.524 4	3.01
		NDCG	0.350 6	0.365 0	4.11
	5	MAP	0.213 0	0.220 2	3.36
		MRR	0.479 1	0.493 2	2.95
		NDCG	0.307 6	0.317 8	3.31

本节实验进一步研究了评分误差权重对推荐结果的影响,在数据集 ml-100k、ml-ls、ml-10m2k 中,top-N 取 5 的情况下,各个评估指标的实验结果如图 2 所示。作为对比,图 2 将同时画出同样条件下 fsSLIM 的评估指标情况,由于不存在评分误差权重,所以 fsSLIM 在所有图中都是直线,每一条直线代表的数据集为距离其最近的折线所代表的数据集。

由图 2 可看出,RCfsSLIM 的推荐效果普遍都比 fsSLIM 要好。同时当评分误差权重偏大或偏小时,推荐效果都有所下降。在数据集 ml-100k 中,评分误差权重取 0.4 时推荐效果普遍最好;在数据集 ml-ls 中,评分误差权重取 0.8 时推荐效果普遍最好;在数据集 ml-10m2k 中,评分误差权重取 0.5 时推荐效果普遍最好。

另外,在评分误差权重很小的情况下,部分数据集的 RCfsSLIM 的推荐效果比 fsSLIM 差,说明 RCfsSLIM 在不对受到用户评分上下文影响的评分进行完全拟合的同时,也不可完全舍去这一部分评分,其在一定程度上有助于模型构建物品相似度矩阵。

3.4.4 在 RCfsSLIM 模型的两个阶段中进行加权处理的有效性

本文 RCfsSLIM 模型在第一个阶段即特征选择阶段中,对评分进行加权处理以计算物品的最近邻;在第二个阶段即模型训练阶段中,使用坐标下降法进行目标函数的优化时,对评分误差进行加权处理以减少算法模型对受到上下文影响的评分的拟合,从而减少该部分评分对物品相似度的影响。

本节实验进一步研究了仅在第一阶段或者仅在第二阶段进行加权处理的推荐效果,在数据集 ml-100k、ml-ls、ml-10m2k 中,top-N 取 3 的情况下,各个评估指标的实验结果如图 3 所示。其中,fsSLIM 模型表示第一个阶段和第二个阶段都不进行加权处理的情况;fsSLIM-weightknn 模型表示仅在第一个阶段即特征选择阶段进行加权处理的情况;fsSLIM-weighterror 模型表示仅在第二个阶段对评分误差进行加权处理的情况;RCfsSLIM 模型即本文方法则表示同时在两个阶段进行加权处理的情况。

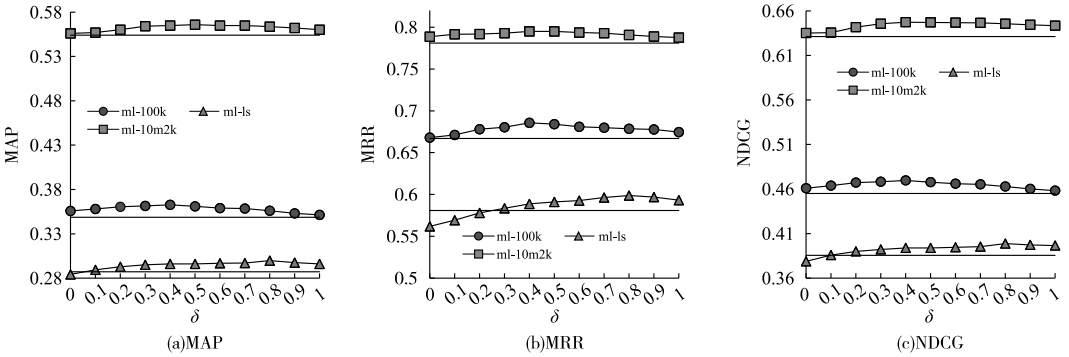


图 2　评分误差权重的影响
Fig. 2　Impact of rating error weight

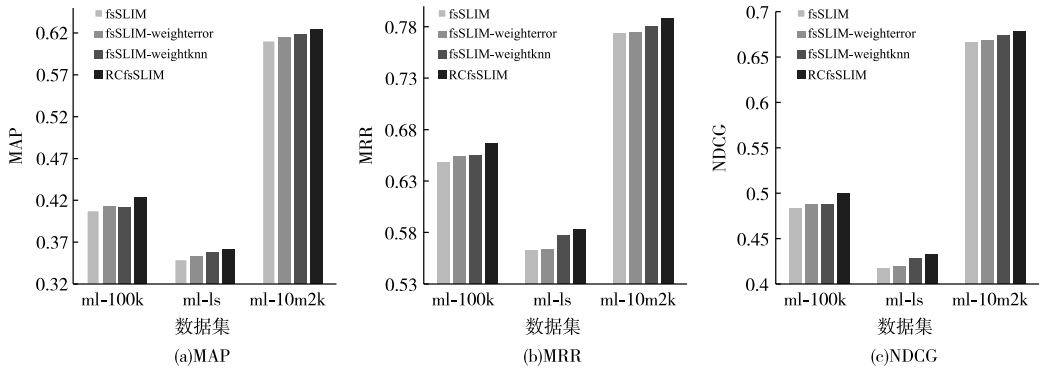


图 3　进行加权处理的有效性
Fig. 3　Effectiveness of weighted

由图 3 可看出,无论是仅在第一个阶段对评分进行加权处理,还是仅在第二个阶段对评分误差进行加权处理,其推荐效果都比不进行加权处理的情况要好,这是因为在第一个阶段即特征选择阶段中对评分进行加权处理可以更好地找出某个物品的相似物品,在第二个阶段即模型训练阶段中对评分误差进行加权处理可以减少算法模型对受到上下文影响的评分的拟合,从而减少该部分评分对物品相似度的影响以提高物品相似度矩阵的准确性。同时,本文提出的在第一个阶段和第二个阶段都进行加权处理的 RCfsSLIM 模型表现出最好的推荐效果。

4　结束语

在一些使用用户评分计算物品相似度的推荐算法中,受到评分上下文影响的用户主观评分并不能客观地反映出物品之间的相似度,从而影响推荐的性能。本文在 fsSLIM 模型的基础上提出了融合评分上下文和物品相似度的 RCfsSLIM 推荐模型。RCfsSLIM 模型首先使用加权评分计算物品相似度的方法进行特征选择,在进一步使用坐标下降法优化目标函数时,对受到用户评分上下文影响的评分误差进行加权处理,使 RCfsSLIM 推荐模型减小对这些评分的拟合以减小该评分对计算物品相似度的影响,从而提高模型训练得到的物品相似度矩阵的准确性。实验结果验证了本文算法的可行性和有效性。

未来将进行以下两方面的工作:a)当前设置某一评分的权重的做法是将用户评分按时间排序后,考虑该评分的先前一个评分的情况,未来工作将综合考虑该评分的先前多个评分的情况,以另一种角度分析评分上下文;b)由于评分权重矩阵的值为人工提前设置,下一步将研究通过算法模型自动学习出各个评分的权重值以提高推荐性能。

参考文献:

[1] Xu Kerui, Yang Jingxuan, Xu Jun, *et al*. Adapting user preference to online feedback in multi-round conversational recommendation[C]// Proc of the 14th ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2021: 364-372.

[2] Srivastava R, Palshikar G K, Chaurasia S, *et al*. What's next? A recommendation system for industrial training[J]. Data Science and Engineering, 2018, 3(3): 232-247.

[3] Liu Luyao, Du Xingzhong, Zhu Lei, *et al*. Learning discrete hashing towards efficient fashion recommendation[J]. Data Science and Engineering, 2018, 3(4): 307-322.

[4] 陈碧毅, 黄玲, 王昌栋, 等. 融合显式反馈与隐式反馈的协同过滤推荐算法[J]. 软件学报, 2020, 31(3): 794-805. (Chen Biyi, Huang Ling, Wang Changdong, *et al*. Explicit and implicit feedback based collaborative filtering algorithm[J]. Journal of Software, 2020, 31(3): 794-805.)

[5] Askari B, Szlichta J, Salehi-Abari A. Variational autoencoders for top-*k* recommendation with implicit feedback[C]// Proc of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2021: 2061-2065.

[6] Lin Guanyu, Liang Feng, Pan Weike, *et al*. FedRec: federated recommendation with explicit feedback[J]. IEEE Intelligent Systems, 2021, 36(5): 21-30.

[7] Jadidinejad A H, Macdonald C, Ounis I. Unifying explicit and implicit feedback for rating prediction and ranking recommendation tasks[C]// Proc of ACM SIGIR International Conference on Theory of Information Retrieval. New York: ACM Press, 2019: 149-156.

[8] Lu Hongyu, Zhang Min, Ma Shaoping. Between clicks and satisfaction: study on multi-phase user preferences and satisfaction for online news reading[C]// Proc of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2018: 435-444.

[9] Ning Xia, Karypis G. SLIM: sparse linear methods for top-*n* recommender systems[C]// Proc of the 11th IEEE International Confe-

rence on Data Mining. Piscataway, NJ: IEEE Press, 2011: 497-506.
[10] Deshpande M, Karypis G. Item-based top-*N* recommendation algorithms[J]. ACM Trans on Information Systems, 2004, 22(1): 143-177.
[11] Choi M, Kim J, Lee J, *et al*. Session-aware linear item-item models for session-based recommendation [C]// Proc of Web Conference. New York: ACM Press, 2021: 2186-2197.
[12] Linden G, Smith B, York J. Amazon. com recommendations: item-to-item collaborative filtering[J]. IEEE Internet Computing, 2003, 7(1): 76-80.
[13] Wang Shoujin, Cao Longbing, Wang Yan. A survey on session-based recommender systems [J]. ACM Computing Surveys, 2022, 54: 1-38.
[14] 包玄,陈红梅,肖清. 融入时间的兴趣点协同推荐算法[J]. 计算机应用,2021,41(8):2406-2411. (Bao Xuan, Chen Hongmei, Xiao Qing. Time-incorporated point-of-interest collaborative recommendation algorithm [J]. Journal of Computer Applications, 2021, 41(8): 2406-2411.)
[15] Tang Jiaxi, Wang Ke. Personalized top-*N* sequential recommendation via convolutional sequence embedding [C]// Proc of the 11th ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2018: 565-573.
[16] Kang Wangcheng, McAuley J. Self-attentive sequential recommendation [C]// Proc of the 18th IEEE International Conference on Data Mining. Piscataway, NJ: IEEE Press, 2018: 197-206.
[17] Sun Fei, Liu Jun, Wu Jian, *et al*. BERT4Rec: sequential recommendation with bidirectional encoder representations from transformer [C]// Proc of the 28th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2019: 1441-1450.
[18] Ozgobek O, Gulla J A, Erdur R C. A survey on challenges and methods in news recommendation [C]// Proc of the 10th International Conference on Web Information Systems and Technologies. 2014: 278-285.
[19] Van B J, Goethals B. High-dimensional sparse embeddings for collaborative filtering [C]// Proc of Web Conference. New York: ACM Press, 2021: 575-581.
[20] Bhattacharya M, Barapatre A. Query as context for item-to-item recommendation [C]// Proc of the 14th ACM Conference on Recommender Systems. New York: ACM Press, 2020: 575-576.
[21] Kabbur S, Ning X, Karypis G. FISM: factored item similarity models for top-*N* recommender systems [C]// Proc of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2013: 659-667.
[22] Lin Jing, Pan Weike, Ming Zhong. FISSA: fusing item similarity models with self-attention networks for sequential recommendation [C]// Proc of the 14th ACM Conference on Recommender Systems. New York: ACM Press, 2020: 130-139.
[23] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent[J]. Journal of Statistical Software, 2010, 33(1): 1-22.
[24] Pan Rong, Zhou Yunhong, Cao Bin, *et al*. One-class collaborative filtering [C]// Proc of the 8th IEEE International Conference on Data Mining. Piscataway, NJ: IEEE Press, 2008: 502-511.
[25] Hu Yifan, Koren Y, Volinsky C. Collaborative filtering for implicit feedback datasets [C]// Proc of the 8th IEEE International Conference on Data Mining. Piscataway, NJ: IEEE Press, 2008: 263-272.
[26] Tay Y, Anh T L, Hui S C. Latent relational metric learning via memory-based attention for collaborative ranking [C]// Proc of the 26th International Conference on World Wide Web. New York: ACM Press, 2018: 729-739.
[27] Song Weiping, Shi Chence, Xiao Zhiping, *et al*. AutoInt: automatic feature interaction learning via self-attentive neural networks [C]// Proc of the 28th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2019: 1161-1170.

(上接第 3031 页)

[11] Singh S S, Srivastva D, Kumar A, *et al*. FLP-ID: fuzzy-based link prediction in multiplex social networks using information diffusion perspective[J]. Knowledge-Based Systems, 2022, 248(19): article ID 108821.
[12] Wu Xiaomin, Wu Jianshe, Li Yafeng, *et al*. Link prediction of time-evolving network based on node ranking [J]. Knowledge-Based Systems, 2020, 195(11): article ID 105740.
[13] Coskun M, Koyuturk M. Node similarity-based graph convolution for link prediction in biological networks[J]. Bioinformatics, 2021, 37(23): 4501-4508.
[14] Kumar A, Singh S S, Singh K, *et al*. Level-2 node clustering coefficient-based link prediction[J]. Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies, 2019, 49(7): 2762-2779.
[15] Yu Chuanming, Zhao Xiaoli, An Lu, *et al*. Similarity-based link prediction in social networks: a path and node combined approach[J]. Journal of Information Science, 2017, 43(5): 683-695.
[16] Chi Kuo, Yin Guisheng, Dong Yuxin, *et al*. Link prediction in dynamic networks based on the attraction force between nodes [J]. Knowledge-Based Systems, 2019, 181(1): article ID 104792.
[17] Caloghirou Y, Giotopoulos I, Kontolaimou A, *et al*. Industry-university knowledge flows and product innovation: how do knowledge stocks and crisis matter? [J]. Research Policy, 2021, 50(3): article ID 104195.
[18] Burt R S. Structural holes: the social structure of competition[M]. Cambridge, MA: Harvard University Press, 1995.
[19] Castela B, Ferreira F, Ferreira J, *et al*. Assessing the innovation capability of small-and medium-sized enterprises using a non-parametric and integrative approach[J]. Management Decision, 2018, 56(6): 1365-1383.
[20] Tao Xinmin, Guo Wenjie, Ren Chao, *et al*. Density peak clustering using global and local consistency adjustable manifold distance[J]. Information Sciences, 2021, 577(8): 769-804.
[21] Lorrain F, White H C. Structural equivalence of individuals in social networks[J]. Social Networks, 1977, 1(1): 67-98.
[22] Lyu Linyuan, Jin Cihang, Zhou Tao. Similarity index based on local paths for link prediction of complex networks[J]. Physical Review E, 2009, 80(4): article ID 046122.
[23] Yao Yabing, Zhang Ruisheng, Yang Fan, *et al*. Link prediction based on local weighted paths for complex networks[J]. International Journal of Modern Physics C, 2017, 28(4): article ID 1750053.
[24] Xia Jingjing, Ling Guang, Fan Qingju, *et al*. Evidential link prediction method based on the importance of high-order path index[J]. Modern Physics Letters B, 2021, 35(33): article ID 2150487.
[25] Zhang Xue, Zhao Chengli, Wang Xiaojie, *et al*. Identifying missing and spurious interactions in directed networks[J]. International Journal of Distributed Sensor Networks, 2015, 8491: 470-481.
[26] Fawcett T. An introduction to ROC analysis[J]. Pattern Recognition Letters, 2005, 27(8): 861-874.

中国计算机学会(CCF)会刊
计算领域高质量科技期刊分级目录T2级(排名第一)期刊
CCF推荐B类中文科技期刊
中国科学引文数据库(CSCD)来源期刊
中文核心期刊
中国科技核心期刊
瑞典开放存取期刊目录(DOAJ)收录期刊
美国剑桥科学文摘(CSA)收录期刊
美国乌利希期刊指南(UPD)收录期刊
日本科学技术振兴机构数据库(JST)收录期刊
中国国际影响力优秀学术期刊
中国WJCI-Top20 期刊
重庆市出版专项资金资助期刊
重庆市“十四五”重点出版物项目资助期刊(领军型学术类)

计算机科学
1974年1月创刊（月刊）
第50卷 第6A期
2023年6月15日出版

COMPUTER SCIENCE
Started in January, 1974(Monthly)
Vol.50, No.6A
Published on June 15th, 2023

主管/主办：重庆西南信息有限公司（原科学技术部西南信息中心）
编辑出版：《计算机科学》编辑部
地　　址：重庆市渝北区洪湖西路18号
邮　　编：401121
电　　话：(023)63500828
电子邮箱：jsjkx12@163.com
社　　长：陈　伟
总　　编：徐书令
编辑部主任：李亚辉
印　　刷：重庆紫石东南印务有限公司
总 发 行：重庆市邮政局
国外发行：中国国际图书贸易总公司（北京399信箱）
邮发代号：78-68
国外代号：6210-MO

Sponsor: Chongqing Southwest Information Co., Ltd. (Southwest Information Center of Ministry of Science and Technology)
Publisher: Editorial Office of *Computer Science*
Address: 18# Honghu West Rd., Yubei, Chongqing, China
Post Code: 401121
Tel: +86-23-63500828
E-mail: jsjkx12@163.com
Proprieter: CHEN Wei
Editor-in-chief: XU Shuling
Editorial Director: LI Yahui
Printer: Chongqing Purplestone Southeast Printing Co.,Ltd.
Distributor: Post Office, Chongqing
Foreign Distributor: China International Book Trading Corporation (P.O. Box 399, Beijing)
Post Issue Number: 78-68
Abroad Code: 6210-MO

国内外公开发行　文章电子版开放获取
http://www.jsjkx.com

国际标准连续出版物号：ISSN 1002-137X
国内统一连续出版物号：CN 50-1075/TP

ISSN 1002-137X
9 771002 137230 99

计算机科学 COMPUTER SCIENCE 2023 Vol.50 No.6A

CS Computer Science
中国计算机学会会刊

ISSN 1002-137X
CN 50-1075/TP
CODEN JKIEBK

计算机科学

2023年 第50卷 第6A期
Vol.50 No.6A 2023



本期封面：人工智能可解释性：发展与应用
王冬丽，杨珊，欧阳万里，李抱朴，周彦

计算机科学

（JISUANJI　KEXUE）

第 50 卷　第 6A 期　2023 年 6 月

目　次

人工智能

220200120 知识增强的自然语言生成研究综述　梁明轩　王　石　朱俊武 等

220600212 人工智能可解释性:发展与应用　王冬丽　杨　珊　欧阳万里 等

220200119 命名实体识别任务综述　高　翔　王　石　朱俊武 等

220400077 方面级情感分析综述　李　阳　王　石　朱俊武 等

220600194 基于经验动态建模的因果检测研究综述　曹旨昊　牟少敏　屈洪春

220600013 命题逻辑中的 L-型冗余性质　刘凌荣　陈树伟　姜世攀

220200112 基于 DCNN 和 GLU 的武器领域实体关系抽取方法　李　晗　侯守璐　佟　强 等

220700127 基于 RRT 的无人机动态航路规划算法　顾子侣　刘　宇　孙文邦 等

220300131 一种新的全局优化算法:碳循环算法　杨　达　罗　亮　郑　龙

220300279 基于 prompt 和知识增强的方面级情感分析　李　阳　唐积强　朱俊武 等

220300141 使用语义解析构建面向分布式 SCADA 系统的自然语言接口　王　涛　郭武士　邓　健 等

220700039 基于弱化图卷积网络的文本分类　黄玉娇　陈铭凯　郑　媛 等

220700031 基于 MADDPG 的无人机群空中拦截作战决策研究　蔺向阳　邢清华　邢怀玺

220500038 基于改进 B-RRT* 算法的移动机器人路径规划　喻九阳　张德安　戴耀南 等

220700153 基于知识增强的命名实体识别方法研究　高　翔　唐积强　朱俊武 等

220200068 概率布尔控制网络的可观性分析　樊卓优

211100070 基于不定可扩展性概念再探对角线方法　段天龙

220300232 基于迁移学习的跨对象手语手势识别方法　王天然　王　琦　王青山

220400164 基于多种强调机制的深度点云网络改进研究　刘　慧　田帅华

220300018 基于情感-主题-讽刺混合模型的讽刺检测研究　付　月　史　伟

220300202 一种新的代价敏感 SVDD 二类分类方法　吴崇明　王晓丹　赵振冲

220700060 基于仿生蛇形机器人的智能蜿蜒运动控制方法　杜洪健　郭　鹏　肖文煜 等

220600230 基于因果推断的图注意力网络　张　涛　程毅飞　孙欣煦

220800211 改进 MFCC 和并行混合模型的语音情感识别　崔　琳　崔晨露　刘政伟 等

220600128 基于多重有向加权图与卷积神经网络的脑电情感识别　罗瑞奇　严金林　胡新荣 等

220100111 联合语义分割和深度估计的多任务学习研究　罗会兰　叶　桔

220400029 基于多图特征聚合的小样本学习方法　曾　武　毛国君

220800189 基于句间信息的图注意力卷积网络的文档级关系抽取　段建勇　杨　潇　王　昊 等

220300275 基于多尺度原型分层匹配的小样本分割方法　孙开伟　刘　虎　冉　雪 等

210300179 基于知识蒸馏的抽取式自动摘要模型 赵江江 王 洋 许楹楹 等
220800192 基于 Doc2Vec 增强特征的长文本主题聚类研究 陈 洁
220700026 多策略离散人工蜂群算法设计 FIR 低通数字滤波器 邵 鹏
220300262 基于双流结构缩放和多重注意力机制的轻量级脑电情感识别方法 雷 颖 刘 峰
220700079 基于多元约束 Petri 网的水利测绘无人机路径规划 姚 喜 陈衍德

图像处理&多媒体技术

220300094 心电信号降噪算法研究综述 侯彦荣 刘瑞霞 舒明雷 等
220400214 基于激光雷达点云的 3D 目标检测方法综述 秦 静 王伟滨 邹启杰 等
220500001 基于数据融合的半监督高分遥感影像语义分割 顾宇航 郝 洁 陈 兵
220500168 虚拟现实在体育仿真中的应用综述:2003 年以来的新进展 纪庆革 陈浩东 何穗申 等
220100158 基于改进的 ResNeXt 网络结构的遥感图像分类 杨 星 宋玲玲 王时绘
220500040 基于孪生注意力网络的建设用地遥感影像变化检测 李 滔 王海瑞
220400171 基于改进 PL-VIO 算法的机器人视觉惯性光学测量法 王海芳 李鸣飞 李广宇 等
220300265 基于亮度校正和融合通道先验的内窥镜图像增强算法 安子恒 徐 超 冯 博 等
220700016 基于稀疏编码非局部注意力对偶网络的病理图像超分辨率重建 梁美彦 张 宇 梁建安 等
220600047 综合应用 Faster R-CNN 和 U-net 的心脏 MRI 图像分割 韩俊玲 李 博 康晓东 等
220300224 基于改进 Cascade R-CNN 的布匹瑕疵检测算法 白明丽 王明文
220300092 基于独立注意力机制的图像检索算法 张舜尧 李华旺 张永合 等
220600129 基于多尺度注意力机制的两阶段文物图像修复方法 刘浩威 姚镜池 刘 波 等
220200092 基于主动学习和 U-Net++分割的芯片封装空洞率的研究 齐选龙 陈弘扬 赵文兵 等
220600182 结合多聚焦融合和 DSGEF 双阶段网络重建太阳斑点图 金亚辉 蒋慕蓉 李福海 等
221100046 基于度量正则化的红外与可见光跨模态行人重识别算法 吴含笑 赵倩倩 朱建清 等
220600240 结构化混合注意力网络的图像超分辨率重建 喻九阳 张德安 戴耀南 等
220500230 回归收敛缩放混合的深度迭代复合缩放 CNN 目标检测算法 王国刚 吴 艳 刘一博
220500104 结合残差与自注意力机制的图卷积小样本图像分类网络 李 凡 贾东立 姚昱旻 等
220600114 基于球簇聚类的超像素分割迭代算法 刘 垚 官礼和
220600142 基于 CT 图像语义的 COVID-19 实例分割与分类网络 柏正尧 樊圣澜 陆倩杰 等
220400023 基于注意力机制最大化重叠的单目标跟踪算法 孙开伟 王支浩 刘 虎 等
220300214 基于注意力机制和可变形卷积的路面裂缝检测 隆 涛 董安国 刘来君
220100205 基于字形感知和注意力归一化的字体迁移 吕文锐 普园媛 赵征鹏 等
220400273 基于 SegFormer 的超声影像图像分割 杨靖怡 李 芳 康晓东 等
220500005 基于改进 YOLOv5 的电动车头盔佩戴检测算法 谢溥轩 崔金荣 赵 敏
220800127 基于交替训练及预训练的低资源泰语语音合成 蔡浩然 杨 鉴 杨 琳 等
220200096 基于区域注意力机制和多尺度特征融合的输电线路螺栓缺陷检测 吴刘宸 张 辉 刘嘉轩 等
211200205 基于运动对比度增强的人群运动分割方法 张新峰 倪启立 陈舒涵 等
220700137 基于边缘优化和全局建模的多路径语义分割 陈乔松 张 羽 蒲 柳 等
220200084 基于渐进式注意力金字塔的行人重识别方法 张帅宇 彭 力 戴菲菲
220200066 基于深度学习的摩托车车道实时检测 万海波 蒋 磊 王 晓

220200063 CT 影像阶段化目标检测方法研究 王笑天 李 博 康晓东 等
220300237 注意力特征融合的孪生网络目标跟踪方法 罗会兰 龙 珺 梁苗苗
220700006 基于改进 Yolov4-tiny 的轻量型目标检测算法 窦 智 胡晨光 梁竞一 等
220700030 基于双门控-残差特征融合的跨模态图文检索 张昌凡 马远远 刘建华 等
220700141 基于机器视觉的驾驶员视野盲区安全预警方法研究 王 巍 白 龙 马欢畅 等
220100138 面向目标识别的特征融合模糊模型及其应用 阮 旺 郝国生 王 霞 等
220400199 联合人体姿态估计和多目标跟踪的跨数据集学习 曾泽华 罗会兰
220200108 基于多编码器的多模态 MRI 脑肿瘤分割 戴天虹 宋洁绮
220600171 基于对比学习的低光照图像增强 吴巨峰 赵训刚 周 强 等

大数据&数据科学

220600241 改进的森林优化特征选择算法在信用评估中的应用 黄宇航 宋 友 王宝会
220600256 GDLIN:一种利用梯度下降的学习索引 陈珊珊 高 隽 马振禹
220500290 基于嵌套集合模型的时态层次数据管理方法 杨振凯 曹一冰 赵鑫科 等
220700094 基于多模态特征融合的时间序列异常检测 张国华 燕雪峰 关东海
221000080 区块链架构下医疗数据共享的三方演化博弈研究 杨 健 王开选
220200015 基于多目标粒子群优化的属性网络局部社区检测算法 周志强 朱 焱
220200042 面向交通流量预测的时空 Graph-CoordAttention 网络 刘建松 康 雁 李 浩 等
220100161 基于改进 CNN-BP 的多波束声纳高程数据预测研究 熊豪杰 魏 怡
211200233 基于主动重心的青年高血压患者心肺运动时序数据增强 黄昉菀 卢举鸿 於志勇
220900137 文本细粒度情绪识别方法与应用综述 王希雅 张 宁 程 馨
220100237 基于图 OLAP 的学术网络分析 杨 恒 朱 焱
220600266 基于动态时空神经网络的城市交通流量预测方法 孟祥福 许睿航
220400127 基于 DBSCAN 的动态邻域密度聚类算法 张 朋 李小林 王李妍
220100066 一种时序情感记忆可约束可解释的序列推荐方法 郑 麟 林艺璇 周东霖 等
220300084 基于决策树改进深度交叉网络的推荐模型 柯海萍 毛宜军 古万荣

网络&通信

220300088 面向能耗优化和负载均衡的边缘服务器放置研究 付 雄 房 磊 王俊昌
211200229 基于线性规划松弛的移动边缘计算卸载模型 雷雪梅 刘 丽 王 倩
220700024 Pauli 噪声环境下任意二粒子受控短距离隐形传态 向生建
220400185 SWIPT-MISO 动态能量消耗模型下能效规划 徐晨阳 薛 亮 王金龙 等
220300272 基于 CEEMDAN-ConvLSTM 组合模型的云计算负载预测方法 赵 鹏 周建涛 赵大明
220600179 基于 Wi-Fi 6 的无线列车通信网络性能研究 杨少龙 朱国胜 庞兴龙 等
220200151 基于深度学习的超高频标签识别系统 余加宝 姚俊梅 谢瑞桃 等
220100166 WSN 中基于改进蝴蝶优化算法的簇首选取算法 杨诗雨 赵 冰 彭 月
220500055 无源 UHF RFID 标签缺失定位特征估计算法 赵 阳 李凌云 赵晓霞 等
220100082 中继选择和队列稳定动态能量优化策略 陈 澈 郑艺峰 杨敬民 等

信息安全

220400200 区块链共识算法综述 谭朋柳 王润庶 曾文豪 等

220300107 RSA 算法在网络数据传输中的研究进展 王鑫淼 孙婷婷 马晶军

220600176 ATT&CK 框架下基于事件序列关联的网络高级威胁检测系统 张宇翔 韩久江 刘 建 等

220300220 基于 CPN 的供应链合约的形式化验证 郑 红 钱诗慧 刘泽润 等

220200121 基于压缩感知和超混沌系统的图像压缩加密方法 潘 涛 佟晓筠 张 淼 等

220100255 基于 IR2Vec 模型的跨架构密码算法识别 赵晨霞 舒 辉 沙子涵

220300064 基于可验证随机函数的实用拜占庭共识算法 黄保华 彭 丽 赵伟宏 等

220300100 基于可跟踪环签名的拜占庭容错共识算法 涂 俊 贾东立 王 津

220400122 基于相似度的 DGA 域名检测方法 孙海栋 刘万平 黄 东

220100157 基于多变量公钥密码系统的环机密交易协议 洪 璇 袁梦玲

220100189 带有衰减效应和遗忘机制的微博网络谣言传播模型 王 韩 刘万平 卢 玲

220400151 基于 PRF-RFECV 特征优选的 GA-LightGBM 的网络安全态势评估 任高科 莫秀良

220700158 一种基于区块链的身份鉴证与授权机制 林飞龙 岳跃栋 郑建辉 等

220400293 基于改进模糊综合评价法的电力监控系统网络可靠性分析 邴英澳 王文婷 孙圣泽 等

220300205 基于分布式集群节点的宕机重启恢复算法 潘 路 罗 涛 牛新征

220600172 基于 FlexUDA 模型的 SQL 注入检测研究 王清宇 王海瑞 朱贵富 等

220500032 基于图像数据耦合识别的输电线路安全风险评估方法 徐昌前 王 东 苏 峰 等

230200188 一种基于自适应加权的鲁棒联邦学习算法 张连福 谭作文

220300034 一种新的基于量子小波变换的图像水印算法 苏永红 夏 婷 王绪梅 等

软件&交叉

220600221 开源软件中社区文档应用与维护的实证研究 张 禹 王 哲 李志星 等

210800242 基于多目特征交叉的服务推荐算法 高文斌 王 睿 祖家琛 等

220500099 有界偏序自动机的同步算法 王志喜 蒋贵德

221100052 新能源汽车供应链的关键风险节点识别方法 杨小博 高海伟 刘天越 等

220500046 基于机器学习的高空电磁脉冲环境快速计算方法 王锦锦 程引会 聂 鑫 等

220400108 基于 Coq 的逆矩阵运算的形式化 沈 楠 陈 钢

220300284 基于流线距离聚类的海洋数据向量场可视化 王 朕 杨政威 高顺起 等

220300023 一种基于网格的水利测绘无人机跟踪方法 姚 喜

220300218 基于多传感器的室内建图导航系统的设计 刘佳伟 杜 欣 范方朝 等

220300114 基于界面相似度的 Android 仿冒应用检测研究 付 雄 聂晓晗 王俊昌

220400128 基于核鲁棒流形非负矩阵分解和融合特征的柴油机故障诊断 刘弘毅 王 瑞 吴贯锋 等

220400216 基于熵权-AHP 与云模型的国产 BIM 建模软件多维度评价研究 赵雪锋 侯 笑 孙 哲 等

220800215 基于混合关键系统的容器调度架构设计 邓广宏 张棋恒

220500056 基于多特征融合的 GRU-LSTM 大学生就业动态预测 张 剑 张 烨

220700258 天然气智能调度计算机平台系统的设计与实现 邓盛男 罗太宇 黄静才 等

220300113 基于博弈论的再制造企业产品回收模型研究 蔡 然 黄鹏鹏

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220300084

基于决策树改进深度交叉网络的推荐模型

柯海萍　毛宜军　古万荣
华南农业大学数学与信息学院　广州 510642
(kehaiping@stu.scau.edu.cn)

摘　要　特征挖掘是推荐算法模型中学习用户与物品之间交互行为的关键步骤,对提升推荐模型的准确度具有重要意义。现有的特征挖掘模型中,线性逻辑回归模型虽然简便,能够达到很好的拟合效果,但其泛化能力较弱,且模型对特征参数量的需求较大。深度交叉网络能够有效实现对特征的交叉提取,但其对数据特征的表征能力仍然不足。因此,文中引入多重残差结构与交叉编码思想,提出了一种基于决策树的方法来改进深度交叉网络的推荐模型。首先基于GBDT算法设计构建强化特征的树结构,加强模型对潜在特征的深度挖掘;其次对模型嵌入层的输入参数维度进行扩增优化;最后对改进的深度交叉网络推荐模型进行推荐预测。该设计不仅可以克服现有模型在泛化能力上的局限性,还能在保持特征参数量精简的同时令其表征能力有所加强,进而有效挖掘用户的隐藏关联,提高推荐的准确度。基于公测数据集的实验结果表明,所提出的模型预测效果比现有的特征交互方法更优。

关键词:特征挖掘;特征交叉;强化特征;决策树;推荐模型
中图法分类号　TP301.6

Recommendation Model Based on Decision Tree and Improved Deep & Cross Network

KE Haiping,MAO Yijun and GU Wanrong
College of Mathematics and Informatics,South China Agricultural University,Guangzhou 510642,China

Abstract　Feature mining is a key step to learn the interaction between users and items in the recommendation algorithm model, which is of great significance to improve the accuracy of the recommendation model. Among the existing feature mining models, although the linear logistic regression model is simple and can achieve good fitting effect,its generalization ability is weak,and the model has a large demand for feature parameters. Deep & Cross network can effectively realize the cross extraction of features, but its representation ability of data features is still insufficient. Therefore,by introducing the idea of multiple residual structure and cross coding,an improved recommendation model of Deep & Cross network based on decision tree is proposed. Firstly,it designs a tree structure based on GBDT algorithm to construct enhanced features,which strengthens the deep mining of the model on potential features. Secondly,the input parameter dimension of the embedded layer of the model is amplified and optimized. Finally,the improved Deep & Cross network recommendation model is used for recommendation prediction. This design can not only break the limitations of existing models in generalization ability,but also keep the feature parameters simple and strengthen their representation ability,so as to effectively mine the hidden associations of users and improve the accuracy of recommendation. Experimental results based on the public test data set show that the prediction effect of the proposed model is better than the exis-ting feature interaction methods.

Keywords　Feature mining,Feature crossover,Enhanced feature,Decision tree,Recommendation model

1　引言

推荐系统已经在电子商务、新闻阅读、视频平台、电影网站、个性化广告投放等领域被各供应商广泛使用[1-3]。传统的推荐方法利用用户的偏好或物品之间的相似度来预测推荐序列[4-5],这些推荐方法的有效性已被YouTube和Amazon证实[6]。传统的协同过滤方法[7-8]和矩阵分解在进行推荐时,仅利用了用户和物品之间存在的显性二元交互信息,而忽略了用户与物品所隐藏的特征交叉关系,这将导致预测的结果序列不够精准[9]。

特征交叉组合[10]是一种处理数据特征的方式,该处理方法增加了特征的维度,使模型能够对特征进行更有效的训练。

基金项目:全国统计科学研究项目(2020LY018);广东省哲学社会科学规划项目(GD19CGL34);中山大学广东省计算科学重点实验室开放基金项目(2021010);广东省自然科学基金面上项目(2022A1515011489);广州市智慧农业重点实验室项目(201902010081)
This work was supported by the National Statistical Science Research Project of China(2020LY018),Philosophy and Social Science Planning Project of Guangdong Province(GD19CGL34),Open Fund Project of Guangdong Key Laboratory of Computing Science,Sun Yat-sen University (2021010),General Program of Guangdong Natural Science Foundation(2022A1515011489) and Guangzhou Key Laboratory of Intelligent Agriculture Project(201902010081).
通信作者:古万荣(guwanrong@scau.edu.cn)

如图 1 中模型对数据特征的处理部分所示，模型对输入的特征进行笛卡尔积来实现特征的组合，组合后得到一组新的特征点。数据特征在相互组合的过程中，可以挖掘到不同特征之间所隐藏的特征，该隐式特征成为用户的一种潜在兴趣点[11-12]。如图 1 所示，以挖掘用户 2 的点击兴趣为例，模型从用户的历史行为数据中对用户曾经选择的物品特征进行笛卡尔积，得到的新特征将指向不同领域的新物品。该新物品与单一的历史选择物品之间存在的直接关联性较弱，但不同历史物品的交集与该新物品的关联性程度却增加了[12]，因此，特征交叉实现了模型对用户隐式特征的挖掘。

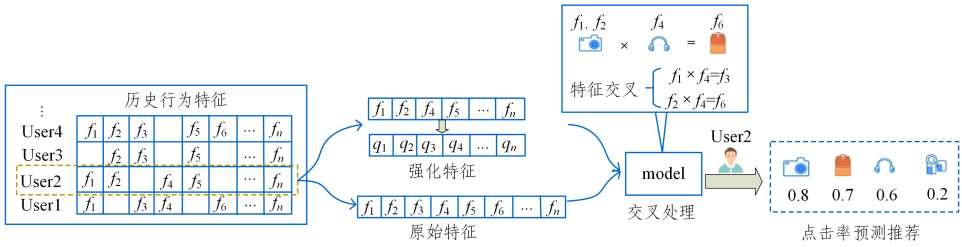


图 1 TDCN 模型工作原理示意图

Fig. 1 Schematic diagram of working principle of TDCN model

为了探讨如何有效挖掘隐藏的特征以提高推荐序列的准确度，基于点击率的推荐模型成为许多学者研究的新方向[13-15]。基于点击率的推荐是利用用户的历史行为数据特征来挖掘用户的潜在交互特征，然后训练预测模型以得到用户可能点击的商品的点击分数，根据点击分数的大小排列得到推荐系统的预测序列。因此，Facebook[16]提出了梯度提升决策树(Gradient Boosting Decision Tree，GBDT)结合逻辑回归(Logistic Regression，LR)的模型。该模型通过自动筛选和组合特征的方式来预测序列的点击率，被广泛应用到点击率(Click-through Rate，CTR)预测中。但是 LR 模型的表达能力有限，且泛化能力较弱，得到的推荐序列有些片面。因子分解机[17](Factorization Machines，FM)模型也能够自动组合特征并学习其交互关系，但是高阶 FM 会导致模型对数据集的参数量需求较大。因此，有效提升模型的特征交叉能力成为特征工程的重要部分[18]。

深度学习是特征交叉有效提高特征表征能力的重要方法。在当下的工业推荐系统中，深度神经网络[19](Deep Neural Network，DNN)模型作为主流模型被许多研究者当作模型改进的基线。该模型能够很好地捕捉特征之间的复杂交互，与线性 LR 模型相比，其充分展现了先进的性能，但隐式的特征组合模式却降低了模型对特征学习的效率。深度因子分解机[20](DeepFM)模型采用低阶和高阶特征组合的方式，有效提高了模型的特征表示能力。深度交叉网络[21](Deep & Cross Network，DCN)模型则提出了一种新颖的特征交叉网络结构，它采用特征交叉相乘和特征交叉的非线性变化方法并行处理的方式来预测推荐序列的点击率，以简洁的网络结构完成了高效的特征交叉工程。

本文针对现有模型的局限性提出了一种基于决策树改进深度交叉网络(Improved Deep & Cross Network Based on Decision Tree，TDCN)的结构，该结构旨在结合多重残差和特征交叉提取的方法来强化数据特征的表征能力，并提高推荐序列点击率的准确性。如图 1 模型的工作原理示意图所示，模型通过构建强化特征，对原特征数据进行了更复杂的交互挖掘，优化了特征在模型中的执行权重比，使模型能够有效利用高相关性的重要特征值，从而提高点击率预测的准确率，以确保推荐系统得到的推荐序列与用户的点击兴趣更相近。本文的主要贡献如下：

(1)通过决策树算法来改进 DCN 模型，既降低了决策树的复杂性，又提高了 DCN 模型的交叉学习能力，能够很好地解决模型过拟合的问题，并且有效提高了模型的泛化能力。

(2)采用决策树算法使模型具有更好的可解释性和鲁棒性，决策树能够自动组合多个特征，较好地处理特征之间的交互，并且能够自动挖掘特征之间的高阶关系，因而降低了对数据集的特征参数量的要求。

(3)使用决策回归树的方法来构建强化特征，增强了模型处理数据特征的能力，提高了模型挖掘潜在特征的效率，从而使模型达到增强特征表征能力的目的，提高模型预测点击率的准确度。

2 相关工作

随着当下公测数据集规模和维数的增加，研究者们对特征工程进行了更加深入的探讨，CTR 预测便在推荐系统中发挥了重要的作用[22-23]。CTR 预测问题类似于常用的推荐问题，并且可以转化为普通的二分类问题。对 CTR 预测模型进行分析，线性 LR[24-25]模型、GBDT[16]算法和 FM[17]模型均被广泛应用到各领域的推荐中，但这些模型更适合用于处理非交互的数据特征场景，因此对数据特征的参数量需求较大。

因子分解机 FM[26]模型主要是对稀疏特征进行映射处理，从而学习特征之间的交互关系。由于 FM 的线性嵌入方法在稀疏数据预测中最有效，因此研究者们以 FM 作为研究基线提出了许多 FM 的变体来提高模型的预测能力。例如 FFM[22]模型(Field-aware Factorization Machines)为每一维特征增加了特征域(Field)，令每个特定的特征域与特征之间形成一对多的关系进行训练，使得模型对特征的提取更精准；FmFM[27]模型(Field-matrixed Factorization Machines)则使用 Field matrix 来建模两个特征之间的交互关系，并计算特征交互的重要性，由于其将交互矩阵进行了相应的扩展，因此模型在保持较高的计算效率的同时还实现了更好的预测效果。

在 CTR 预测与推荐系统的深度学习模型中，DNN[19]模型以多个全连接层的方式完成对特征的交叉和组合，但这种直接的方式在特征的信息表征上欠缺针对性。由于全连接层没有对不同特征域进行交叉处理，而实际问题中的特征之间包含了大量高价值的潜在信息，这促使当下研究者需要探索更多有针对性的深度学习网络结构来表征这些特征交叉信息。

Wide & Deep[6]模型融合了浅层 Wide 网络与深层 Deep 网络结构，其利用浅层网络对交叉特征的记忆能力以及深层网络对特征的泛化能力进行组合训练，但在模型联合训练时，Wide 网络部分处理的特征较小，特征的正确选择直接决定了

模型的成功与否，因此模型的泛化能力有待优化。DCN[21]模型则使用了 Cross 网络来替换 Wide & Deep 模型中的 Wide 网络部分，并通过叠加输入特征的方式来实现特征的自动学习，在数据特征的选择和交互上都具有明显的优势。而 Lian 等[28]提出了一种压缩交互网络(Compressed Interaction Network，CIN)的极深因子分解机(xDeepFM)模型来实现向量级的特征交互，CIN 以显式与隐式结合的模式自动建模高阶特征交互。Meng 等[29]则利用基于梯度的搜索策略进行探索，提出了一种能够自动挖掘有效特征的强交互通用模型。

针对现有模型存在的不足，本文提出了一种基于决策树改进深度交叉网络(TDCN)的推荐模型来强化数据特征的表征能力，利用改进的决策提升树对数据特征进行自动组合筛选，并利用深度交叉网络组件对数据作特征交叉处理，简化了模型对数据特征的参数量要求，并在 DCN 模型良好的特征交叉提取能力下进一步加强了特征之间的交互力度，对用户的特征进行了更深层次的挖掘提取，增强了模型的泛化能力，提高了推荐序列的准确度。

3 模型介绍

本文基于决策树提出了一种改进的深度交叉网络(TDCN)模型来提升推荐序列的点击率准确度。该模型利用决策树对特征组合提取的优势以及深度交叉网络对特征交叉的挖掘能力，以改进的决策树算法来构建强化特征，从多方面增强模型对特征的表征能力，从而提高推荐序列的点击率预测准确度。模型的整体结构如图 2 所示，模型由去噪强化特征构建的树模块和 DCN 组件两部分组成。

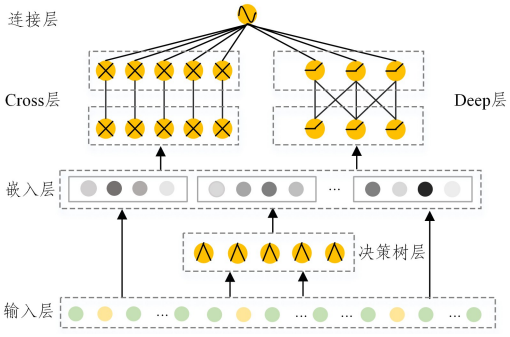


图 2 TDCN 架构图

Fig. 2 Architecture diagram of TDCN

本文提出的 TDCN 模型对 DCN 模型的嵌入层进行了特征强化的改进。由图 2 可知，TDCN 网络在决策树模块构建了强化特征，并将强化特征向量输出到嵌入层，该层将所有特征向量进行叠加，得到对应的特征叠加向量将作为模型的初始输入特征向量。

3.1 去噪强化特征构建的树模块

基于树的方法[30-32]被广泛应用于分类、排序和各种机器学习的模型中。现有的 CTR 预测模型中，DeepFM[20]，DCN[21]，xDeepFM[28]和 AutoInt[33]均能够实现特征交叉挖掘。这些模型利用了独立并行的子网络来进行特征的交互，挖掘信息之间的潜在特征。模型对特征交互的挖掘能捕捉显性特征和隐性特征之间的相关性，但庞大的特征参数会削弱主要特征的权重值[34]，从而降低模型的预测准确率。Guo 等[35]便尝试以分布式的方式来训练大规模参数量的嵌入表，提出了一种基于混合缓存的 CTR 模型分布式训练系统。

为了克服这一局限性，本文提出了一种基于决策树实现的去噪强化特征构建方案，该方案由本文提出的决策树模块实现，用于提取特征数据中的重要信息，并根据提取的特征来构建强化特征，以达到对有效特征去噪的目的。与现有的大多数模型相比，该模块的设计与原模型的特征输入部分有所不同，强化特征的加入使模型在特征嵌入层增加了输入特征向量的参数量，利用改进的决策树模块增加了重要特征数据的权重占比，从而提高了模型的预测准确度。

具体来说，在 TDCN 模型中，决策树模块构建的去噪强化特征增加了模型处理特征向量的维度，从而使模型着重挖掘输入数据中强相关性的隐藏特征信息。受 GBDT+LR[16]模型的 GBDT 算法设计以及 RLNF[36]模型(Reinforcement Learning based Noise Filtering)对特征降噪滤波的启发，本文提出了一个改进的决策树模块，以用数据特征中的重要相关性特征来构建去噪强化特征。

在模块的构建中，决策树模块使用矩阵 $\boldsymbol{X}\in R^{n\times d}$ 作为存储器，其中 n 为样本数量，d 为矩阵的维数，即特征的数量。模块首先对输入的特征矩阵进行重要性计算，剔除弱关联的噪声特征，同时提取相关性程度较大的特征作为待输入的样本数据。现假设输入的样本数据为 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\cdots,\boldsymbol{x}_n)^{\mathrm{T}}$，$x_i$ 指第 i 个样本数据向量，使用 $T_M(x)$ 表示第 M 棵决策树的处理函数，$F_M(x)$ 为包含前 M 棵决策树的提升树加法模型。因此，决策提升树模块可以用迭代公式表达为：

$$F_M(x)=F_{M-1}(x)+T_M(x) \tag{1}$$

模块在更新的过程中，使用平方损失函数对每轮迭代的结果与真实值 y 计算残差值 r，则第 m 轮的损失函数计算式如下：

$$r=y-F_{M-1}(x) \tag{2}$$

$$L(y,F_M(x))=(y-F_M(x))^2=(r-T_M(x))^2 \tag{3}$$

然后可以根据每轮决策得到的叶子节点数值进行模块最终的结果向量计算，其计算公式如下：

$$\tilde{x}=\sum_{i=1}^{d}w_i x_i \tag{4}$$

其中，$\tilde{x}$ 为决策提升树模型的最终输出结果，即强化特征值，w_i 为第 i 个特征的权重值。最后强化特征向量作为决策树模块的输出结果，在 TDCN 模型的下一网络层被进一步利用。

总而言之，改进的决策树模块是通过构建去噪强化特征的方式来加强模型对特征的挖掘重心，最大化有效特征的挖掘强度，避免模型的预测准确率随特征参数量的增加而降低。因此，去噪强化特征构建的树模块对模型预测准确率有重要的提升作用。

3.2 DCN 组件

DCN 模型是一个并行的网络结构，它由 Cross 网络和 Deep 网络两个独立的子模块组成，分别以不同的处理方法对输入的特征向量进行特征交叉挖掘，这两个子模块的表达式分别为：

$$x_{l+1}=x_0 x_l^{\mathrm{T}} w_{c,l}+b_{c,l}+x_l \tag{5}$$

$$h_{l+1}=\mathrm{ReLu}(W_{d,l}^{T}h_l+b_{d,l}) \tag{6}$$

其中，x_l 和 h_l 表示 Cross 网络和 Deep 网络第 l 层的输出值，而 $w_{c,l}$ 和 $b_{c,l}$ 分别为 Cross 网络第 l 层的权重参数和偏置参数；同理，$w_{d,l}$ 和 $b_{d,l}$ 为 Deep 网络第 l 层的权重参数和偏置参数，另外，$x_0=h_0$，即两个子网络的初始输入值相同。

在 TDCN 模型中，DCN 模块的初始输入值在嵌入叠加层

进行处理。特别地，由于经过 One-Hot 处理后的类别特征具有稀疏性，于是对该类特征向量进行等维度的嵌入化处理，计算公式为：

$$x_{e,i}=w_{e,i}x_{\text{sparse},i} \tag{7}$$

其中，$x_{\text{sparse},i}$ 为类别特征向量，$w_{e,i}$ 为嵌入化转换的权重。因此该网络层的主要功能是将输入模型的特征数据完成向量化的转换，并叠加对应的嵌入向量，得到 DCN 模块的初始输入值如下：

$$x_0=[x_{e,1}^{\mathrm{T}},x_{e,2}^{\mathrm{T}},x_{e,3}^{\mathrm{T}},\cdots,x_{e,k}^{\mathrm{T}},x_{\text{dense}}^{\mathrm{T}},x_{\text{enhance}}^{\mathrm{T}}] \tag{8}$$

其中，$x_{e,k}^{\mathrm{T}}$ 中包含 k 个类别特征的嵌入向量，$x_{\text{dense}}^{\mathrm{T}}$ 表示数值特征向量，$x_{\text{enhance}}^{\mathrm{T}}$ 表示强化特征向量。

最后，两个子网络层的输出在连接层叠加，并对叠加得到的向量进行结果预测。连接层的表达公式如下：

$$p=\text{sigmoid}(w^{\mathrm{T}}[x_L,h_L]+b) \tag{9}$$

其中，p 为模型的预测结果，sigmoid 为模型预测使用的激活函数，而 w 和 b 分别为连接层的权重和偏置参数。

综上所述，除了嵌入层和决策树层，TDCN 模型中其他各子网络层的工作原理与现有的 DCN 模型相同。如图 2 所示，本文提出的模型在改进的决策树模块中构建强化特征，在 DCN 组件的 Cross 子网络对数据特征进行线性变换处理以及在 Deep 子网络对数据特征进行非线性变换处理。

此外，本文以 AUC 评估指标作为模型性能评估的主要评估标准，并使用 Adam 算法对模型进行优化。AUC 指标的公式表达如下：

$$AUC=\frac{\sum_{positives}k-\frac{n_{\text{pos}}(n_{\text{pos}}-1)}{2}}{n_{\text{pos}}n_{\text{neg}}} \tag{10}$$

其中，$\sum_{positives}k$ 为序列正样本排列的固定值，n_{pos} 为序列的正样本数值，n_{neg} 为序列的负样本数值。得到的 AUC 数值越大代表模型的的效果越好。

本文提出的 TDCN 模型的最大优势在于构建了强化特征，巧妙利用了决策树与深度交叉网络的互补优势，并有效缓解了决策树容易过拟合的局限性。由不同数据集在 TDCN 模型与 DCN 模型中的对比结果，可以看出 TDCN 模型的点击率预测准确率有所提高。

4 实验

4.1 实验数据集

本文实验采用 Criteo[1)] 数据集和 Avazu[2)] 数据集这两个公测数据集分别对模型进行性能评估。

(1)Criteo 数据集：该数据被广泛使用在广告点击率相关的实验中，它包含有 4500 万用户的点击日志，总共由 13 个连续特征(I1－I13)和 26 个类别特征(C1－C26)组成，并且每个类别都是高基数的级别，本文实验中所使用的 sample Criteo 是对 Criteo 数据集的采样，有 100 万个样本。

(2)Avazu 数据集：该数据集是由美国互联网广告公司提供的用户 11 天的在线点击历史记录，其中包含用户所点击内容的标志、领域、类别等共 23 个特征数据。该数据集被广泛应用于点击率预测的模型中进行训练，让模型学习高数量级的特征交叉计算，从而提高系统的整体预测性能。

这两种数据集通常作为模型预测广告点击率的训练数据，由于数据集中包含了大量用户在不同平台中对广告推荐的点击情况，十分适合用于类比推荐系统中预测的序列被用户感兴趣并点击浏览的可能性。两种数据集还有一个共同的特点，即它们同时存在连续特征和类别特征，这些特征的交互训练，在推荐系统的预测序列中能够很好地体现推荐序列的被点击率，由此来提高推荐系统对序列点击率预测的性能。

4.2 评估指标

实验中将两个数据集预处理之后分为训练集和测试集。训练集的作用是为了不断调整模型的参数，直至模型的训练结果达到最优，并取其最优结果的参数，而评估模型的性能优劣则主要在测试集中进行。本文实验使用交叉熵(Logloss)和 AUC 作为模型的评估指标，并将 AUC 评估指标作为本文实验的主要参考指标。其中，Logloss 评估指标的分数值越小，代表模型的性能越佳。AUC 值则介于 0 到 1 之间，越接近 1 代表模型分类器的性能越好。

4.3 对比基线

实验的对比基线选用现有的 AFM[37]，NFM[38]，PNN[39] 和 DCN[3)] 模型。

(1)AFM：该模型通过引入 Attention 机制来学习不同组合特征的权重，增强模型预测的可解释性。

(2)NFM：该模型提出一种特征交叉池化层的结构来建模多个特征之间的深层交互关系，从而提高组合模型对高低阶特征交互的能力。

(3)PNN：该模型采用乘法运算来加强 DNN 网络结构对特征交叉的表征能力，以缓解交叉特征表达不充分的问题。

(4)DCN：该模型引入了交叉网络结构，能够自动构造有限高阶的交叉特征，并学习对应权重，从而以轻量级的特征参数量完成高效的特征交叉处理。

本文提出的 TDCN 模型中，Cross 层、Deep 层和连接层采用了与 DCN 模型相同的交叉处理结构，主要的优势在于增加了决策树层以及对嵌入层做了改进。TDCN 模型的决策树层对各特征进行了特征重要性和关联性的挖掘，得到对应的重要特征来构建强化特征，DCN 组件则对扩充维度后的数据特征进行对应的交叉挖掘，进而实现模型整体对潜在隐式特征的准确挖掘，以提高模型预测推荐序列的准确度。实验中用两种公测数据集来比较模型的点击率预测效果，以各评估指标的变化曲线图来直观地对比模型之间的性能效果。

4.4 实验结果与分析

实验结果部分主要展现了本文模型的部分运行效果以及对模型整体预测性能的评估对比结果。为了更直观地了解决策树层对特征的筛选，我们在实验一对算法的特征筛选结果进行了排序。此外，模型整体的预测性能评估结果决定着该模型的可行性，我们分别通过实验二和实验三来说明模型的评估结果，而评估的对比结果将进一步验证模型的性能优越性。在模型评估部分的可视化对比实验中，为了更直观地

1) http://labs.criteo.com/2014/02/kaggle-display-advertising-challenge-dataset/

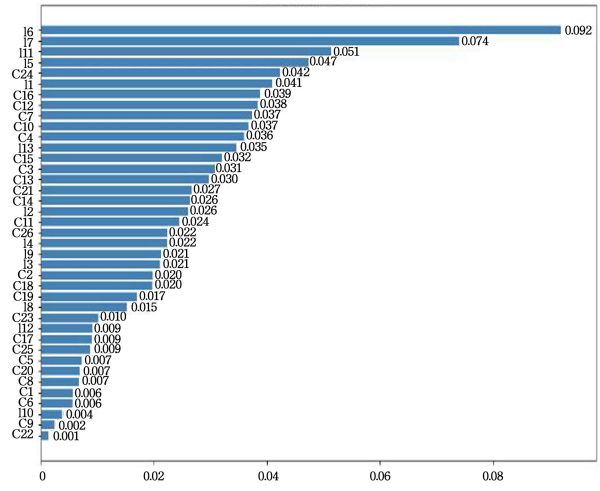
2) https://www.kaggle.com/c/avazu-ctr-prediction

3) https://github.com/princewen/tensorflow_practice/blob/master/recommendation/Basic-DCN-Demo/DCN.py

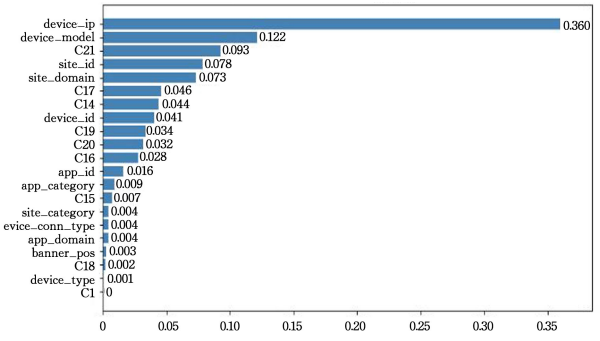
对比和验证模型的预测性能，实验仅在 DCN 模型和 TDCN 模型上使用两种公测数据集分别进行预测评估。

4.4.1 实验一：特征筛选

特征相关度是影响实验预测结果的因素之一，为了更好地理解实验数据集中每个特征对预测结果的影响，该部分的实验在特征筛选时对每个特征都进行了重要性分数的评测，并以重要性程度进行排序。如图 3 展示的结果为数据集中各特征与预测结果的相关性关系，对其归一化处理得到每个特征的重要性分数，其中，横坐标为每个特征的重要性分数值，纵坐标按分数的高低来排序所有的特征。对特征进行相关性分析能够有效剔除对点击结果影响较弱的噪声特征，提取出最具代表性的特征来构建有利于模型挖掘的强化特征，从而使模型在整体训练过程中，增加对有效特征的学习权重，更好地提高点击率预测的准确度。



(a)Criteo 数据集的特征分析实验结果



(b)Avazu 数据集的特征分析实验结果

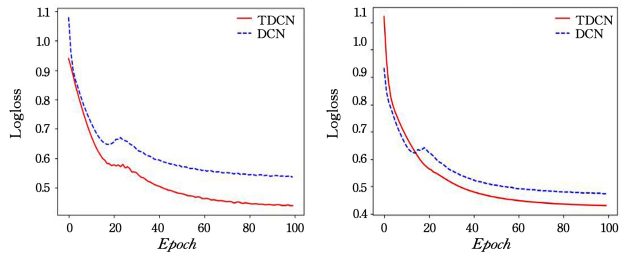
图 3 特征筛选实验结果

Fig. 3 Experimental results of characteristicscreening

从实验结果可以看出，各数据集的特征排序中前 10 个特征的重要性占总特征重要性的一半以上，而后 10 个特征只占总特征重要性的 5%左右。因此，模型在决策树层对数据特征进行特征筛选时，将后 10 个特征看成噪声特征并剔除，选用其余特征来构建强化特征，因而得到的强化特征具有特征的强代表性。

4.4.2 实验二：模型的 Logloss 评估

Logloss 评估是二分类模型最常用的评估指标之一，它对预测概率具有较好的敏感性，主要用来衡量模型预测的概率值与真实值之间的差异性，即得到的分数越小越好。在这一部分的实验中，模型所得到的 Logloss 评估对比结果如图 4 所示，实验中分别对数据集进行了 100 次的迭代，得到 Logloss 的分数收敛趋势图。Logloss 评估分数取值为 0～1，随着实验的迭代次数增加而逐渐趋于一个较稳定的分数，该稳定值为模型的最终评估分数。



(a)Criteo 数据集的 Logloss 评估结果 (b)Avazu 数据集的 Logloss 评估结果

图 4 模型的 Logloss 评估结果

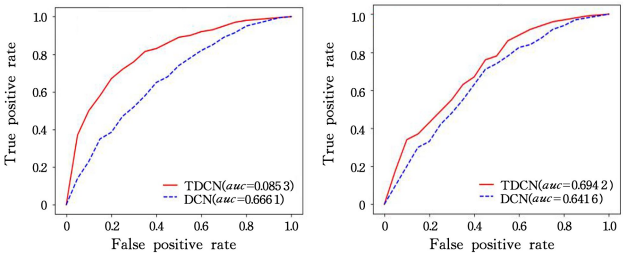
Fig. 4 Logloss evaluation results of models

从图 4 的 Logloss 评估结果可以得到，随着实验的 Epoch 值不断增加，两种模型的损失函数值都开始逐渐收敛，TDCN 模型的评估值在收敛之后明显比 DCN 模型的评估值更低。另外，TDCN 模型的评估曲线变化趋势比 DCN 模型更加平滑，尤其在 Avazu 数据集中表现得最为明显，说明特征参数量越少的数据集在 TDCN 模型中训练后的预测效果比现有的 DCN 模型的预测效果更优，并且 TDCN 模型的鲁棒性更强。

4.4.3 实验三：模型的 AUC 评估

AUC 指标只能用于衡量二分类模型的优劣，该值必须大于 0.5，并且越接近 1 代表模型的分类器效果越好。该实验的主要目的是评估 DCN 模型和 TDCN 模型中的二分类预测效果如何。模型在两个公测数据集上的评估结果对比如图 5 所示，横坐标代表样本预测的错误率，纵坐标代表样本预测的正确率，其 ROC 曲线越凸向左上角方向说明模型的二分类预测性能越优。实验得到的 AUC 值是 ROC 曲线下的面积，即对随机给定的一个样本，模型预测结果的正确率比错误率更大的可能性数值，该值越高说明模型的二分类预测效果越好。

由图 5 中两种模型所得到的 ROC 曲线对比图可知，TDCN 模型的 AUC 值比 DCN 模型更高，说明 TDCN 模型的二分类预测效果更好。此外，模型在 Criteo 数据集中得到的评估结果要比 Avazu 数据集更加明显，说明 TDCN 模型在处理高阶复杂的数据量特征时，特征的表征能力更强，性能表现效果更优。



(a)Criteo 数据集的 AUC 评估结果 (b)Avazu 数据集的 AUC 评估结果

图 5 模型的 AUC 评估结果

Fig. 5 AUC evaluation results of models

4.4.4 模型整体性能的比较

实验中分别采用 Logloss 和 AUC 两个评估指标来衡量不同模型的性能，表 1 总结了 TDCN 模型和 4 种对比基线分别在不同数据集上的实验评估结果。

表 1 模型的实验评估结果

Table 1 Experimental evaluation results of models

method	Criteo		Avazu	
	Logloss	*AUC*	*Logloss*	*AUC*
AFM	0.4554	0.7953	0.5325	0.6247
NFM	0.4537	0.7968	0.5268	0.6316
PNN	0.4521	0.7987	0.4543	0.6830
DCN	0.5298	0.6651	0.4618	0.6501
TDCN	**0.4390**	**0.8099**	**0.4488**	**0.7011**
Impr./%	17.1	**21.8**	2.8	**7.8**

由表 1 可知，本文提出的 TDCN 模型比所有对比基线方法的评估效果都更优。以 AUC 评估指标为主要标准，可以得到 TDCN 模型的评估结果在 Criteo 数据集和 Avazu 数据集中相比 DCN 模型的评估结果分别提升了 21.8%和 7.8%，可以进一步验证得到本文主要探讨的结论：

(1)本文对模型的特征数据集进行重要性筛选，并采用改进的决策树算法来构建去噪强化特征，使模型对结果序列的预测优于 DCN 模型的预测结果，验证了本文提出的 TDCN 模型经过特征强化的加持后具有更好的可实施性。

(2)本文提出的 TDCN 模型使推荐序列的点击预测结果准确率得到进一步的提升，从而验证了构建强化特征有利于提高模型对用户隐式偏好的学习与深度挖掘，能够更加有效地预测用户的点击偏好。

结束语 隐式交叉关联特征信息的深度挖掘对提高推荐模型的点击率预测准确度具有重要意义。本文对并行结构的深度模型(DCN)进行了探讨，并针对传统方法的不足，提出了一种基于决策树改进深度交叉网络(TDCN)的推荐模型来加强对隐式特征信息的挖掘，它由去噪强化特征构建的树模块和 DCN 组件模块组成。TDCN 模型在决策树层中对数据集中的相关性特征进行了筛选，使用改进的决策提升树算法对重要特征进行强化特征的构建，以去噪的强化特征来增加模型处理有效特征的数据维度，使模型处理的数据具有明确的特征表达性，从而很好地加强 TDCN 模型对用户特征和物品交互特征的挖掘。

本文基于两个公测数据集对模型进行了不同的性能评估实验对比，以证明 TDCN 模型的有效性和兼容性。最终的实验结果表明，本文提出的 TDCN 模型在多种评估方法中的结果均比 DCN 模型更好，因此，TDCN 模型的推荐准确度比 DCN 模型更高，也证实了有效的特征处理能够较好地提高模型预测的准确度。此外，测试结果也表明了 TDCN 模型在特征工程中具有广泛的应用可扩展性，尤其是一些追求用户点击能够直接带来效益的应用平台。

本文在当前的工作中没有考虑以不同的算法来区分不同类型的属性交互。在未来的研究中，将从属性的类型为切入，并结合上下文信息来强化模型对高阶特征交互的能力，优化模型在各层网络结构对数据特征处理的性能，研究以多级优化的方法来进一步提升推荐的准确度与泛化性能。

致谢 本文在此感谢广州市智慧农业重点实验室(201902010081)的资助。

参 考 文 献

[1] BARROS M, COUTO F M, PATO M, et al. Creating Recommender Systems Datasets in Scientific Fields[C]// Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2021: 4029-4030.

[2] HAO X B, LIU Y D, XIE R B, et al. Adversarial Feature Translation for Multi-domain Recommendation[C]// Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2021: 2964-2973.

[3] GUPTA U, WU C J, WANG X D, et al. The architectural implications of facebook's DNN-based personalized recommendation [C]// Proceedings of the 2020 IEEE International Symposium on High Performance Computer Architecture. IEEE, 2020: 488-501.

[4] DONG M Q, YUAN F, YAO L N, et al. MAMO: Memory-Augmented Meta-Optimization for Cold-start Recommendation [C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 688-697.

[5] HUANG Z, TAO M Y, ZHANG B F. Deep Inclusion Relation-aware Network for User Response Prediction at Fliggy[C]// Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021: 3059-3067.

[6] CHENG H, KOC L, HARMSEN J, et al. Wide & Deep learning for recommender systems[C]// Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. 2016: 7-10.

[7] ZHANG S C. Research on Recommendation Algorithm Based on Collaborative Filtering[C]// Proceedings of the 2nd International Conference on Artificial Intelligence and Information Systems. 2021: 1-4.

[8] GONG L X, WANG J Y. Research on Collaborative Filtering Recommendation Algorithm for Improving User Similarity Calculation[C]// Proceedings of the 2021 International Conference on Control and Intelligent Robotics. 2021: 331-336.

[9] XU J P, WU L F, PANG X L, et al. 2nd International Workshop on Industrial Recommendation Systems[C]// Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021: 4173-4174.

[10] LANG L, ZHU Z L, LIU X Y, et al. Architecture and Operation Adaptive Network for Online Recommendations[C]// Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021: 3139-3149.

[11] GUO H F, CHEN B, TANG R M, et al. An Embedding Learning Framework for Numerical Features in CTR Prediction[C]// Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021: 2910-2918.

[12] LI P, JIANG Z C, QUE M F, et al. Dual Attentive Sequential Learning for Cross-Domain Click-Through Rate Prediction [C]// Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021: 3172-3180.

[13] GUO L Y, JIN J Q, ZHANG H Q, et al. We Know What You Want: An Advertising Strategy Recommender System for Online Advertising[C]// Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021: 2919-2927.

[14] ZHOU G R, ZHU X Q, SONG C R, et al. Deep interest network for click-through rate prediction[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 1059-1068.

[15] WANG P, JIANG Y, XU C, et al. Overview of Content-Based Click-Through Rate Prediction Challenge for Video Recommendation[C]// Proceedings of the 27th ACM International Confe-

rence on Multimedia. 2019：2593-2596.
[16] HE X，PAN J，JIN O，et al. Practical lessons from predicting clicks on ads at facebook[C]// Proceedings of the 8th International Workshop on Data Mining for Online Advertising. 2014：1-9.
[17] CHEN C，ZHANG M，MA W Z，et al. Efficient Non-Sampling Factorization Machines for Optimal Context-Aware Recommendation[C]// Proceedings of The Web Conference 2020. 2020：2400-2410.
[18] LIU B，ZHU C X，LI G L，et al. AutoFIS：Automatic Feature Interaction Selection in Factorization Models for Click-Through Rate Prediction[C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020：2636-2645.
[19] COVINGTON P，ADAMS J，SARGIN E. Deep neural networks for YouTube recommendations[C]// Proceedings of the 10th ACM Conference on Recommender Systems. 2016：191-198.
[20] GUO H，TANG R，YE Y，et al. DeepFM：a factorization-machine based neural network for CTR prediction[C]// Proceedings of the 26th International Joint Conference on Artificial Intelligence. 2017：1725-1731.
[21] WANG R，FU B，FU G，et al. Deep & Cross network for ad click predictions[C]// Proceedings of the ADKDD'17. 2017：1-7.
[22] JUAN Y，ZHUANG Y，CHIN W，et al. Field-aware factorization machines for CTR prediction[C]// Proceedings of the 10th ACM Conference on Recommender Systems. 2016：43-50.
[23] CARLOS M P，CAMILO V，JUAN M M，et al. Leveraging User Embeddings and Text to Improve CTR Predictions With Deep Recommender Systems[C]// Proceedings of the Recommender Systems Challenge 2020. 2020：11-15.
[24] PRAOWPAN T. Identifying key drivers in airline recommendations using logistic regression from web scraping[C]// Proceedings of the 2020 the 3rd International Conference on Computers in Management and Business. 2020：112-116.
[25] REN K，ZHANG W，RONG Y，et al. User response learning for directly optimizing campaign performance in display advertising [C]// Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 2016：679-688.
[26] JIANG G W，WANG H，CHEN J，et al. XLightFM：Extremely Memory-Efficient Factorization Machine[C]// Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021：337-346.
[27] SUN Y，PAN J W，ZHANG A，et al. FM2：Field-matrixed Factorization Machines for Recommender Systems[C]// Proceedings of the Web Conference 2021. 2021：2828-2837.
[28] LIAN J X，ZHOU X H，ZHANG F Z，et al. XDeepFM：Combining Explicit and Implicit Feature Interactions for Recommender Systems[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018：1754-1763.
[29] MENG Z，ZHANG J，LI Y，et al. A general method for automatic discovery of powerful interactions in click-through rate prediction[C]// Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021：1298-1307.
[30] SAKURAI K H，SHIMIZU T K. Actor-based incremental tree data processing for large-scale machine learning applications [C]// Proceedings of the 9th ACM SIGPLAN International Workshop on Programming Based on Actors，Agents，and Decentralized Control. 2019：1-10.
[31] HOSSAIN M，RAFI S，HOSSAIN S. An Optimized Decision Tree based Android Malware Detection Approach using Machine Learning[C]// Proceedings of the 7th International Conference on Networking，Systems and Security. 2020：115-125.
[32] WAN X C，ZHANG H，WANG H，et al. RAT-Resilient All reduce Tree for Distributed Machine Learning[C]// Proceedings of the 4th Asia-Pacific Workshop on Networking. 2020：52-57.
[33] SONG W P，SHI C，XIAO Z P，et al. Autoint：Automatic feature interaction learning via self-attentive neural networks[C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019：1161-1170.
[34] CHEN X，DU Y L，XIA L，et al. Reinforcement Recommendation with User Multi-aspect Preference[C]// Proceedings of the Web Conference 2021. 2021：425-435.
[35] GUO H，GUO W，GAO Y，et al. ScaleFreeCTR：MixCache-based Distributed Training System for CTR Models with Huge Embedding Table[C]// Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021：1269-1278.
[36] ZHAO P，LUO C，ZHOU C，et al. RLNF：Reinforcement Learning based Noise Filtering for Click-Through Rate Prediction [C]// Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021：2268-2272.
[37] XIAO J，YE H，HE X，et al. Attentional factorization machines：Learning the weight of feature interactions via attention networks[J]. arXiv：1708.04617，2017.
[38] HE X，CHUA T S. Neural factorization machines for sparse predictive analytics[C]// Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2017：355-364.
[39] QU Y，CAI H，REN K，et al. Product-based neural networks for user response prediction[C]// Proceedings of the 16th International Conference on Data Mining (ICDM). IEEE，2016：1149-1154.



KE Haiping，born in 1997，postgraduate，is a student member of China Computer Federation. Her main research interests include recommendation system and deep learning.



GU Wanrong，born in 1982，Ph.D，master. His main research interests include search engine，Internet big data analysis and mining，recommendation system and biological information mining.

ISSN 1001-9081
CN 51-1307/TP
CODEN JYIIDU

計算機應用

JOURNAL OF COMPUTER APPLICATIONS

四川省计算机学会　中国科学院成都分院　主办

☆ 中文核心期刊
☆ 两届国家期刊奖提名奖
☆ 全国优秀科技期刊一等奖
☆ 中国计算机学会会刊

微信公众号

2024年7月　第44卷 增刊1
Jul. 2024　Vol.44 Suppl.1

国家期刊奖获奖期刊
中国期刊方阵“双奖”期刊

英国《科学文摘》(SA)收录期刊
俄罗斯《文摘杂志》(AJ)收录期刊
美国《剑桥科学文摘》(CSA)收录期刊
德国《数学文摘》(Zentralblatt MATH)收录期刊
波兰《哥白尼索引》(IC)收录期刊
《日本科学技术振兴机构数据库》(JST)收录期刊

中文核心期刊
中国科技核心期刊
RCCSE中国核心学术期刊
中国科学引文数据库(CSCD)刊源期刊
中国科技论文与引文数据库(CSTPCD)统计源期刊
中国学术期刊综合评价数据库(CAJCED)统计源期刊
中国核心期刊遴选数据库收录期刊
中文科技期刊数据库(VIP)收录期刊
CEPS中文电子期刊服务全文收录期刊

计 算 机 应 用

(Jisuanji Yingyong)

第44卷 第S1期 2024年7月

目 次

人工智能

基于文本知识的多跳推理综述 …… 王昕瑞,陈 涛,孙 涛(1)
基于句法CYK图神经网络的知识增强文本分类 …… 章 巍,陈学奇,韩剑锋,虞小江,吴海燕(11)
基于主动学习和持续学习的同义词挖掘模型 …… 潘桢皓,关东海,袁伟伟,郭 然(18)
基于多特征驱动图注意卷积网络的关系抽取 …… 李航程,钟 勇(24)
大语言模型引导的文本摘要技术与系统 …… 黄君豪,朱锦文,向宗元,李萌坚,毛瑞琛(29)
融合在线检索和量化低秩适配器微调范式的新闻文稿生成 …… 励 琦,刘志强,李 岚,向宗元,毛瑞琛,陈 群(34)
基于MLP和多头自注意力特征融合的双模态情感计算模型 …… 吴俊洁,王佳阳,朱 萍,肖 强(39)
基于RoBERTa和集中注意力机制的营商政策多标签分类 …… 陈昊飏(44)
基于文本挖掘的航空公司旅客行为分析 …… 孙天笑,田 勇,陈锦辉,黄钰翔(49)
基于因果关系的小样本类增量学习 …… 刘冰瑶,刘进锋(54)
自适应策略优化的粒子群优化算法在神经网络架构搜索中的应用 …… 程金芮,金 瑾,张朝龙,孔 超,何 嘉,张 鑫(60)
超轻量级点云分类网络 …… 王豪洋,田茂义,俞家勇(65)
基于3D归一化流的双分支点云异常检测模型 …… 王布凡,张绍兵,成 苗,何 莲(71)

数据科学与技术

基于样本选择的自学习去噪鲁棒模型 …… 王 魁,姚 迅,杨 捷(77)
跨域双活数据库事务处理的动态优化技术 …… 吴昊亮,危建朋,崔欣宁,张先伟,陈 丽,谷 峪(83)
异构体系结构下的混合负载数据库自适应并行技术 …… 吴昊亮,刘永宏,危建朋,陈建宇,万 珊,谷 峪(89)
基于深度学习的高能同步辐射光源数据快速无损压缩算法 …… 张敏行,符世园,高 宇,程耀东,陈 刚(95)

网络空间安全

基于区块链免密钥托管的可追踪可撤销属性基加密方案 …… 郭丽峰,马雪珂,王玲霞,张夏蕾(101)
基于联邦学习和区块链的车联网隐私保护数据共享方案 …… 王杨鹏,牛宪华,熊 玲,陈 鹏,徐 雷(107)
分布式电力交易系统安全威胁与检测 …… 李 达,马军伟,郭庆雷,柏德胜,徐若然(112)
基于盲签名的匿名资格证在电子投票中的应用 …… 付铄雅,崔 喆,李美娟(118)

先进计算

图神经网络在脑网络中的应用研究综述 …… 李文杰,韩秋蕾,孙大文,赵 剑,匡哲君(124)

面向流体力学的物理神经网络综述 …… 田松岩,黄鑫格,段焰辉,陈洪波,陈文秀(133)
肿瘤标志物识别的特征选择综述 …… 韩媛媛(142)
移动边缘计算中基于Actor-Critic深度强化学习的任务调度方法 …… 黄一帆,曾 旺,陈哲毅,于正欣,苗 旺(150)
云数据中心能耗感知的虚拟机优化放置方法 …… 苏苗苗,李志华(156)
基于大数据的复杂系统智能建模方法 …… 郭 呈,史春秀,梅顺峰,张 熠(164)
动态网络局部社区演化过程溯源方法 …… 罗香玉,卢 琰,王 甜(169)
概率故障条件下平衡超立方体的子网络可靠性 …… 冯 凯,高红艳(175)

网络与通信

基于时域划分的网络流量预测方法 …… 夏明山,王 丽(183)
具有无线携能技术的MIMO-NOMA系统 …… 李贵勇,高馨雨,于晓娜(188)

计算机软件技术

高性能CAE软件性能评测平台的设计与应用 …… 李于锋,段博文,肖永浩(193)
京东零售基于ClickHouse的增量刷岗方法 …… 季 健,洪 帅,陈洪健,钱 叶,刘传耀(199)

多媒体计算与计算机仿真

从U-Net到Transformer:深度模型在医学图像分割中的应用综述 …… 张玮智,于 谦,苏金善,孚西旦·居马洪,林 玲(204)
基于改进DFVDFF网络的变焦深度测量 …… 赵涂昊,夏小东,付茂栗,王 宪(223)
不均匀光照下的合作目标图像分割方法 …… 黄 亮,郝颖明(229)
基于逐层扩张卷积的3D牙颌实例分割 …… 唐瑞成,成 苗,何 莲,石向文,郭 静,赵铱民,王胜朝,余 快(235)
合成孔径雷达图像与可见光图像配准方法综述 …… 梅益文,回丙伟,郭鹏程(242)
基于自适应加权融合激光雷达和相机的三维目标检测方法 …… 董钰婷,官 磊(250)
基于FastFlow的复杂纹理图像异常检测 …… 张 果,夏小东,付茂栗,魏 麟(256)
基于Deformable DETR的域自适应目标检测模型在城市街道场景中的应用 …… 杨 浩,陈 斌,李 科(262)
基于大模型微调范式的绘画风格模拟方法 …… 马诗洁,徐华艺,李聪聪,耿卫东,沈华清,李萌坚(268)
基于虚拟现实的火星车路径规划方法 …… 刘震磊,王 丹,朱万安,倪浩添,李月月(273)
基于生成对抗网络的多尺度密集残差雨滴去除网络 …… 刘 越,柴秋月,刘 芳,张佳乐,王天笑(277)
基于动态视觉信息归一化的健身动作分析系统 …… 胡海晴,李建伟,薛 珺,李朋杰(284)
基于改进YOLOv7的高效行人检测方法 …… 冯恒健,韩李涛,张鹏飞,李洪梅(290)
融合卷积和注意力机制的车辆行人检测算法 …… 林卓成,赵尔敦,张政阳,黄 洪(297)
基于生成对抗网络的车牌图像篡改检测数据增广 …… 李来源,霍聪聪,谭舜泉(301)
基于检索任务和改进ResNet的机场跑道异物目标识别方法 …… 刘 广,李海旭,侯 腾,王晓成(309)
基于双线性卷积宽度网络的水稻病虫害识别 …… 孙杨俊,陈 滔,刘志梁(314)
基于YOLOv7-tiny的轻量化海珍品检测算法 …… 陈俊逸,曹立杰,吴 军,罗佳璐,何植仟(319)

前沿与综合应用

基于Transformer编码器和残差网络的信贷违约预测模型 …… 张瑶娜,卓佩妍,刘自金,刘 炜,宋 友(324)
基于改进XGBoost的金融客户投资行为特征选择方法 …… 吴成英,马东方(330)
分时电价下的充电站负荷预测 …… 顾 玮,段 敬,张 栋,郝晓伟,薛泓林,安 毅,段 婕(337)
基于双向长短期记忆和多层级联注意力的配电网态势感知模型 …… 袁新平,原 野,王海燕,唐 铭(343)
面向电力输电场景概略点云的无人机巡航线自动生成 …… 徐景中,莫玉晓,付建红,孙红星(347)
基于动态自适应蚁群优化算法的移动机器人路径规划 …… 聂清彬(351)
基于强化学习的离场飞行程序航迹生成方法 …… 宋 歌,韩鹏飞,罗钰翔(355)
基于时序模型的麻醉期人体生命体征预测 …… 汪倍民,倪献春,姚 宇,陈 皎(363)
基于单核苷酸多态性筛选的品种鉴定方法 …… 郭美萍,古万荣,谢先荣,毛宜军,陈墩金,熊 懿(369)
金沙江白鹤滩水电站生态环境智慧管理平台的应用 …… 张嵩榴,王思敏,宗蔷雯,王云锋,杨少伟,廖 勇(374)

本期责任编辑:夏朝晖 聂卫东 胡 丹 胡 奇 田 星 周 强

文章编号:1001-9081(2024)S1-0369-05 DOI:10.11772/j.issn.1001-9081.2023060887

基于单核苷酸多态性筛选的品种鉴定方法

郭美萍[1],古万荣[1*],谢先荣[1],毛宜军[1],陈墩金[2],熊 懿[1]
(1.华南农业大学 数学与信息学院,广州 510642; 2.广州明领基因科技有限公司,广州 510642)
(*通信作者电子邮箱guwanrong@scau.edu.cn)

摘 要:针对基因序列数据集中的高维小样本和样本不平衡问题,提出基于单核苷酸多态性筛选的品种鉴定方法。首先,利用改进的卡方统计量对预处理后的基因序列数据进行卡方值的计算,通过比较卡方值的大小筛选出单核苷酸多态性(SNP)位点;其次,使用变体的过采样技术Variant-SMOTE(Variant Synthetic Minority Over-sampling TEchnique)平衡样本,以所有类别中的最大样本数量为基准,将剩余的其他少数类样本扩充至基准量,再将得到的矩阵输入K近邻(KNN)训练模型;最后,将该模型应用于10个水稻亚种的基因序列数据中迭代10次,取最高、最低的一组数据和平均值与2020—2023年提出的6种方法进行对比实验。实验结果表明,所提模型可以有效地选取出重要度排名前1%的位点作为SNP位点,即可以有效地从3 139 562个位点中选出31 396个高信息量位点用于品种鉴定,鉴定的准确率达到了85.21%。相较于2020—2023年提出的6种方法,所提模型在水稻亚种鉴定中具有最好的鉴定效果,准确率比最低的方法chi-square+random forest提高了120%,比最高的方法chi-square+SMOTE+KNN和ReliefF+SVM提高了11.01%。所提方法虽然较对比方法更耗时,但在保留高信息量SNP的基础上减少了用于品种鉴定的位点数,降低了品种鉴定的成本。

关键词:卡方统计量;过采样;特征选择;品种鉴定;K近邻
中图分类号:TP181 **文献标志码**:A

Variety identification method based on single nucleotide polymorphism screening

GUO Meiping[1], GU Wanrong[1*], XIE Xianrong[1], MAO Yijun[1], CHEN Dunjin[2], XIONG Yi[1]
(1. *College of Mathematics and Informatics, South China Agricultural University, Guangzhou Guangdong* 510642, *China*;
2. *Minglead Gene Technology Corporation Limited, Guangzhou Guangdong* 510642, *China*)

Abstract: A variety identification method based on single nucleotide polymorphism screening was proposed to address the issues of high-dimensional small samples and sample imbalance in gene sequence datasets. Firstly, an improved chi-square statistic was used to calculate the chi-square values of the preprocessed gene sequence data, and Single Nucleotide Polymorphism (SNP) loci were selected by comparing the chi-square values. Secondly, a variant oversampling technique Variant-SMOTE (Variant Synthetic Minority Over-sampling TEchnique) was used to balance the samples. Based on the maximum sample size of all classes, the remaining minority class samples were expanded to the reference size, and the resulting matrix was input into K-Nearest Neighbors (KNN) to train the model. Finally, the model was applied to the gene sequence data of 10 rice subspecies for 10 iterations, and the highest and lowest set of data and average values were taken for comparative experiments with six methods proposed in 2020-2023. The experimental results show that the proposed model can effectively select the top 1% important loci as SNP, that is, 31 396 high information loci can be effectively selected from 3 139 562 loci for variety identification, with an accuracy rate of 85.21%. Compared with the six methods proposed in 2020-2023, the proposed method has the best identification performance in rice subspecies identification, with an accuracy improvement of 120% compared to the lowest method, chi-square+random forest, and 11.01% compared to the highest methods, chi-square+SMOTE+KNN and ReliefF+SVM. Although the proposed method is more time-consuming than the comparison methods, it reduces the number of loci used for variety identification while retaining high information SNPs, thereby reducing the cost of variety identification.

Key words: chi-square statistic; oversampling; feature selection; variety identification; K-Nearest Neighbors (KNN)

0 引言

品种鉴定在生物遗传信息的挖掘中具有广泛和重要的研究意义和遗传育种应用价值[1],为制定育种策略和实施品种选育奠定良好的基础[2]。在植物品种鉴定中,挖掘已公开品种的单核苷酸多态性(Single Nucleotide Polymorphism, SNP),使用机器学习鉴定候选品种或者后代品种面临两个难点:高维小样本和样本不平衡[3]。为了解决这两个难点,本文提出了一种基于SNP筛选的品种鉴定方法用于水稻亚种鉴定。

水稻是亚洲乃至全世界主要的农作物之一。在过去的几十年里,众多研究者利用水稻性状的表型特征鉴定水稻的籼、粳属性。一般地,籼稻和粳稻在形态、抗逆、抗病性以及种植地理分布等方面都存在明显的差异[2]。由于籼、粳稻的这些差异,正确鉴定杂交水稻后代的籼、粳亚种在水稻的育种策略制定、育种选育和品种遗传信息管理意义上显得尤为重要。前人一般采用最直观的方法,即观察粒形来区分籼稻和粳稻,但这并不准确[2]。在学术研究上,程氏指数法是鉴定稻米籼、粳属性最经典的方法,但该方法至少需要追踪水稻一个生长周期,不仅周期长还受季节性影响[4]。为了解决这些问题,本文从水稻叶子中提取SNP进行分析以鉴定水稻的籼、粳属性,SNP指DNA序列上核苷酸位点呈现碱基转换

收稿日期:2023-07-07;**修回日期**:2023-08-29;**录用日期**:2023-09-05。
基金项目:广东省自然科学基金资助项目(2022A1515011489);云浮市科技计划项目(2022020301)。
作者简介:郭美萍(1995—),女,广西贺州人,硕士研究生,主要研究方向:生物信息挖掘、机器学习; 古万荣(1982—),男,广东梅州人,讲师,博士,主要研究方向:搜索引擎、推荐系统、生物信息挖掘; 谢先荣(1988—),男,江西赣州人,讲师,博士,主要研究方向:水稻育性基因的克隆和功能;毛宜军(1979—),男,湖北公安人,讲师,博士,主要研究方向:基因数据处理、信息安全、人工智能; 陈墩金(1983—),男,广东揭阳人,高级管理师,博士,主要研究方向:基因数据处理; 熊懿(1999—),男,湖南益阳人,硕士研究生,主要研究方向:新闻推荐。

中国知网 https://www.cnki.net

或者颠换[2]。

调研相关文献可知，较少研究使用机器学习方法对农作物SNP数据进行处理和品种鉴定，多数研究都是基于农作物的表型性状：Dong等[5]使用聚类分析将水稻小穗表型相关性状分为四类；王彩红等[6]应用程氏指数和109个简单重复序列标记(Simple Sequence Repeats，SSR)对100份亚洲栽培稻品种进行分类，拟合度较高(73.3%～100%)，表明两种方法在分类上存在正相关性；郑向华等[2]将程氏指数作为基线，使用SNP-index(Single Nucleotide Polymorphism index)方法进行品种鉴定，保证40-SNP的真实性与可靠性。观察法鉴定水稻不准确，而程氏指数鉴定水稻周期长，所以本文方法具有研究意义。

基于SNP数据，通过机器学习区分杂交水稻的籼、粳属性以鉴定水稻品种，为水稻育种选育奠定良好基础，显然时间效率远高于程氏指数法。SNP是同一物种中的个体与个体之间脱氧核苷酸(DeoxyriboNucleic Acid，DNA)序列中单个位置的变异，而同一物种中的所有个体在DNA序列上有99.9%的相似性[7]。所以在许多应用于SNP数据分类的机器学习研究中，超高维度超小样本就成为主要挑战，而维度与样本之间的量极差可能达到甚至超过3个量级。许多公共数据库的SNP数据由于采集时的随机性，通常伴随着数据不平衡的问题，即有的种类样本数很多而其他一些种类样本数很少，导致分类器对样本数多的种类学习得更好，对样本数少的种类更容易分类错误，必然会影响模型的分类准确性。

使用SNP数据通过机器学习进行分类的研究更多应用于医学领域，如：Li等[8]使用k中心点算法(k-center clustering algorithm，k-center)对人类遗传病进行聚类，聚类效果不是很好，也没有对SNP数据进行降维；文献[9]中提出的模型也未对SNP数据进行降维处理，且在支持向量机(Support Vector Machine，SVM)分类器上的分类准确率也只有80.35%。目前有许多研究对SNP数据进行了降维，如Alzubi等[10]融合条件互信息度量(Conditional Mutual Information Metric，CMIM)和支持向量机递归特征消除(Support Vector Machine Recursive Feature Elimination，SVM-RFE)进行降维，提出了一个能够分析复杂疾病的SNP数据并区分健康者与患者的框架，但是它在不同疾病的数据集上分类准确率不稳定。在Alzubi等[10]的研究之前还有Evans[11]的CSS(Chi-Squared Sort)和DS(Difference Sort)，Batnyam等[12]的CBFS(Clearness-Based Feature Selection)、ReliefF、RFS(R-value based Feature Selection)、FSDD(Feature Selection based on Distance Discriminant)等，这6种特征选择方法分别对精神发育迟滞(Mental Retardation，MR)、孤独症谱系障碍(Autism Spectrum Disorder，ASD)、乳腺癌(Breast Cancer，BC)这3个SNP数据集进行了降维，但是在SVM模型上的分类准确率低于CMIM和SVM-RFE融合的方法，同样也存在不同数据集分类准确率不稳定的问题。存在这些问题的原因可能是SNP数据集样本不平衡，Gunasekaran等[13]使用了卷积神经网络(Convolutional Neural Network，CNN)和混合神经网络模型对5种病毒的DNA进行了分类，并使用合成少数类过采样技术平衡数据样本，提高了模型的准确率。由于该模型把数据样本输入CNN训练之前直接截取前6 000个SNP位点，尽管CNN模型对数据有特征选择的功能，但是对于百万数量级的SNP数据集并不友好。为了有效缓解这些问题，本文提出用改进的卡方(chi-square)统计量对数据集降维，随后使用变体合成少数类过采样技术(Variant Synthetic Minority Over-sampling TEchnique，Variant-SMOTE)平衡数据样本，最后输入K近邻(K-Nearest Neighbors，KNN)分类器训练模型。

为了验证方法的可行性，本文首先将收集的10种水稻亚种(Aromatic、Aus、Indica1、Indica2、Indica3、IndicaIntermediate、Intermediate、JaponicaIntermediate、TemperateJaponica、TropicalJaponica)的SNP数据进行清洗；其次利用改进的chi-square进行降维，去除冗余SNP位点，得到一个特征SNP子集；再次利用Variant-SMOTE对子集进行类别平衡，最后利用KNN分类器鉴定水稻亚种。通过分类测试实验，验证了本文方法能够利用相对于原始样本的较少量具有高信息量的SNP将多标签的水稻亚种正确分类。

1 材料和方法

1.1 实验数据与预处理

1.1.1 实验数据

本文研究的数据来源于水稻基因组变异的综合数据库RiceVarMap(http://ricevarmap.ncpgr.cn/)数据库的12个归因基因型染色体，是从水稻叶子中测序的SNP数据。本项目在RiceVarMap的下载页面，从12个归因基因型的染色体(chr01、chr02、chr03、chr04、chr05、chr06、chr07、chr08、chr09、chr10、chr11、chr12)中挑选出1 414个SNP位点缺失值较少(缺失率小于3%)的水稻种质材料，每一个水稻种质中都包含了6 279 124个SNP位点信息(已过滤除掉缺失的位点)，其中两个位点组成了双等位基因，即AA、TT、CC、GG。本文使用plink将SNP数据整合成一个ped文件(merge.ped，该文件占存储空间为16.8 GB)，并将所选取的SNP位点信息存储在map文件中。在搜索品种信息页面中，找到每个水稻种质对应的亚种，保存在sample_classify.txt文件中。1 414个水稻种质从属于10个亚种，样本详细信息如表1所示。

表1 水稻数据集

数据集	样本数	SNP碱基对	亚种名称	亚种样本数
水稻	1 414	3 139 562	Aromatic	69
			Aus	42
			Indica1	56
			Indica2	31
			Indica3	151
			IndicaIntermediate	89
			Intermediate	36
			JaponicaIntermediate	161
			TemperateJaponica	375
			TropicalJaponica	400

1.1.2 数据预处理

预处理包括删除冗余的字符、SNP数据格式的转化和SNP编码。merge.ped中包含1 414个水稻种质的3 139 562对双等位基因，每一个等位基因之间有一个空格符，而该数据是纯合子，所以可以删除双等位基因中的其中一个。同时由于PED文件格式的原因，该文件还包含了一些不利于分类的字符，如“0 0 0 -9”，这些冗余字符都要删掉。处理后将大小为16.8 GB的文件压缩至8.4 GB。为了降低SNP序列的操作和分析难度，本文查找sample_classify.txt中的个体类别信息，将数据再次处理后以fasta格式保存数据文件。fasta文本格式用于记录核酸序列或肽序列，其中的核酸或氨基酸均以单个字母编码呈现，大小写均可。在fasta格式中，一条序列分为两行，第一行是序列的描述信息，以一个大于号(“>”)开头，内容可以随意，但不能有重复，相当于身份识别信息。第二行则是序列本身，使用单个字母表示的一行或多行的核苷酸序列或肽序列。原始数据集和预处理后的数据集如表2~3所示。

表2 原始数据集

ID	其他信息	序列
B001_B001	0 0 0 -9	A A G G T T C C … T T G G
B002_B002	0 0 0 -9	A A G G T T C C … T T G G
⋮	⋮	⋮
W207_W207	0 0 0 -9	A A G G T T C C … T T G G

DNA序列很长，本文数据与处理完后的序列长度高达3 139 562，放入分类器中训练具有挑战性，也不能直接作为分类器的输入，必须将数据进行编码。编码方法对分类精度也有着至关重要的作用，机器学习中有很多种编码方法，在这里本文使用一种常用的编码方式——标签编码。标签编码是将序列中的每一个核苷酸都用唯一的索引值表示，保留序列的位置信息[14]，如图1所

中国知网 https://www.cnki.net

示(A-1、C-2、G-3和T-4)。

表3 预处理后的数据集

ID	序列
B001	AGTC…TG
B002	AGTC…TG
⋮	⋮
W207	AGTC…TG

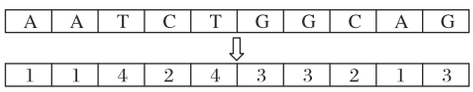


图1 SNP标签编码

1.2 方法

本文首先利用改进的卡方选择找出对区分样本最有用的特征位点,然后,使用Variant-SMOTE算法平衡样本类别,最后使用KNN对样本进行分类。本文方法流程如图2所示。

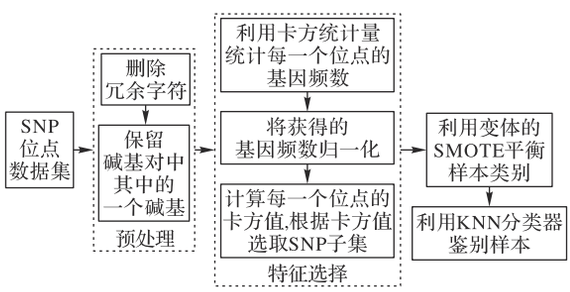


图2 本文方法流程

1.2.1 改进的卡方统计量

特征选择对分类结果影响很大,有必要使用适当的特征选择算法降低数据集的维数。常用的特征选择算法有信息增益(Information Gain, IG)、互信息(Mutual Information, MI)、基尼系数(GIni, GI)、词频(Term Frequency, TF)、文档频率(Document Frequency, DF)、卡方(chi-square)等[15]。其中,卡方统计量适用于离散型数据。本文使用改进的卡方进行选择,改进点在于进行卡方值计算之前加入归一化对基因频数进行标准化。卡方统计量定义为数据的理论分布或假设分布与其实际分布之间的差异的度量,即卡方值的大小取决于观测值与理论值之间的偏离程度,如果卡方值越小,则说明观测值与理论值的偏差越小,反之,二者的偏差越大。当某个特征的两个值完全相等时,计算出来的卡方值等于0,表明观测值完全符合理论值,也就是不相关,即该特征不显著,与样本分类无关,应当删除该特征。根据以上定义,将本文的问题假设定义如下:

H0:假设在样本中每一个位点ATCG的取值是均匀分布的,即ATCG的取值与样本分类无关。

H1:假设在样本中每一个位点ATCG的取值不是均匀分布的,即ATCG的取值与样本分类相关。

结论:根据卡方值判断某个位点是否与样本分类相关,卡方值越大则说明更支持H1假设,即相关性越高,否则H0假设成立,即相关性越低。

根据以上问题描述,本文将样本看作是一个$m \times n$的矩阵,m表示样本数,n表示位点数,$n \gg m$。首先分别统计每一个位点的ATCG的观测频数$f_{a_i}, f_{t_i}, f_{c_i}, f_{g_i}$,然后使用式(1)计算每一个位点的卡方值:

$$\chi^2 = \frac{(f_{a_i} - f_e)^2}{f_e} + \frac{(f_{t_i} - f_e)^2}{f_e} + \frac{(f_{c_i} - f_e)^2}{f_e} + \frac{(f_{g_i} - f_e)^2}{f_e} \tag{1}$$

其中,$i = \{1, 2, \cdots, n\}$,f_e表示每一个位点ATCG的理论频数,$f_e = m/4$。

在卡方计算前加入归一化对基因频数进行标准化。如式(2)所示:

$$\boldsymbol{F}_o = \begin{bmatrix} \frac{f_{a_1} - f_{\min_1}}{f_{\max_1} - f_{\min_1}} & \frac{f_{a_2} - f_{\min_2}}{f_{\max_2} - f_{\min_2}} & \cdots & \frac{f_{a_n} - f_{\min_n}}{f_{\max_n} - f_{\min_n}} \\ \frac{f_{t_1} - f_{\min_1}}{f_{\max_1} - f_{\min_1}} & \frac{f_{t_2} - f_{\min_2}}{f_{\max_2} - f_{\min_2}} & \cdots & \frac{f_{t_n} - f_{\min_n}}{f_{\max_n} - f_{\min_n}} \\ \frac{f_{c_1} - f_{\min_1}}{f_{\max_1} - f_{\min_1}} & \frac{f_{c_2} - f_{\min_2}}{f_{\max_2} - f_{\min_2}} & \cdots & \frac{f_{c_n} - f_{\min_n}}{f_{\max_n} - f_{\min_n}} \\ \frac{f_{g_1} - f_{\min_1}}{f_{\max_1} - f_{\min_1}} & \frac{f_{g_2} - f_{\min_2}}{f_{\max_2} - f_{\min_2}} & \cdots & \frac{f_{g_n} - f_{\min_n}}{f_{\max_n} - f_{\min_n}} \end{bmatrix} \tag{2}$$

其中:$\boldsymbol{F}_o$是一个矩阵,表示标准化的基因频数;$f_{\min_i}$表示第i个位点的最小值,$f_{\min_i} = \min(f_o)$,f_o为每一种碱基的观测频数;$f_{\max_i}$表示第i个位点的最大值,$f_{\max_i} = \max(f_o)$。最后,结合式(1)和式(2)可以得到式(3):

$$\boldsymbol{\chi}^2 = (\boldsymbol{F}_o - \boldsymbol{F}_e)^2 / \boldsymbol{F}_e \tag{3}$$

其中:$\boldsymbol{\chi}^2$是一个矩阵,表示所有位点的卡方值;$\boldsymbol{F}_e$是一个$4 \times n$的矩阵,元素都是f_e。

根据问题假设,和以上描述可以得到2×4列联表,如表4所示。其伪代码如算法1所示。

表4 SNP数据集2×4列联表

碱基	观测值	理论值
A	所在位点A的频数	A的理论频数
T	所在位点T的频数	T的理论频数
C	所在位点C的频数	C的理论频数
G	所在位点G的频数	G的理论频数

算法1 改进的chi-square。

输入 dataset, label, feature。

dataset表示样本集,label表示样本标签,feature表示特征数

输出 SNP。

#SNP表示所筛选的高信息量的位点子集

1)统计样本中每一个位点的A、T、C、G的频数;

2)使用式(2)标准化基因频数;

3)使用式(3)计算每一个位点的卡方值;

4)根据卡方值选择前1%的SNP子集;

5)将SNP子集进行标签编码后返回。

1.2.2 Variant-SMOTE

机器学习中常用于分类的算法,如:逻辑回归(Logistic Regression, LR)、决策树(Decision Tree, DT)、SVM、随机森林(Random Forest, RF)和KNN等,都假设数据样本是平衡的,即不同类别的样本在数量和质量上都是同等的。然而,真实世界中,大多数需要处理的数据样本并不满足这一要求。本文用到的数据集,栽培水稻SNP的少数类样本与多数类样本比接近1:13。若直接使用收集到的不平衡数据集进行分类,那么结果是不太理想的;因此,本文需要解决数据样本的不平衡问题。现有工作中,常用的解决方法大致可以分为三类:特征选择法、数据分布调整法和模型训练算法调整。本文使用数据分布调整法中的合成少数类过采样技术(Synthetic Minority Over-sampling TEchnique, SMOTE),SMOTE的算法原理是根据原始数据样本的分布,通过定义邻域内的几个少数类实例之间进行插值(插值过程如图3所示)而合成新样本[16]。因此,该过程基于距离度量,从训练集中选择同一类的几个最近邻居,在两两连线之间插入一个点生成新的样本。SMOTE算法定义如式(4)所示:

$$\boldsymbol{X}_{(\text{new})} = \boldsymbol{X}_j + (\hat{\boldsymbol{X}} - \boldsymbol{X}_j) \times \delta \tag{4}$$

其中:$\boldsymbol{X}_j$为第j个样本;$\hat{\boldsymbol{X}}$为k个近邻中选择的m个样本;$\delta \in (0, 1)$是一个随机数,表示在两个样本的连线上随机产生一个点作为新合成的少数类样本。

由于传统的SMOTE只适用于二值分类,而本文使用的数据集有10种类别,所以将SMOTE改为变体的SMOTE。该算法基于有放回随机选择的思想,从少类样本中有放回随机选择num个样本,

中国知网 https://www.cnki.net

*num*等于基准数(以所有类别中的最大样本数量为基准值)减去少数类的样本量。然后通过式(4)生成*num*个新样本,将原始样本和生成的样本合在一起,使少数类样本量扩充至基准量。

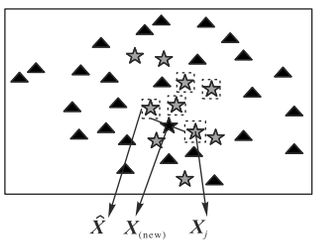


图3 SMOTE插值生成新样本

1.2.3 KNN分类器

KNN是机器学习中最简单、最经典的一种有监督学习分类算法[17],可以有效地解决分类问题。其指导思想是"同类相吸",所谓同类相吸指的是一个样本的周围最邻近的k个样本中大多数属于某一个类别,那么这个样本就属于这个类别。换而言之,根据定类决策寻找最邻近的一个或者几个样本属于哪一个类别来判定待分类的样本的所属类别。具体算法步骤如下:

1)确定k个近邻点;

2)计算未知样本与已知样本之间的距离(本文使用欧氏距离);

3)按照距离递增次序排序;

4)选择与未知样本距离最小的k个样本;

5)确定这k个样本所属类别出现的次数;

6)返回前k个样本出现次数最高的类别作为未知样本的预测分类。

2 实验与结果分析

本文所提出的模型使用Python进行编程,在Windows 11系统上运行,处理器为AMD Ryzen 7 5800H with Radeon Graphics。

2.1 数据集

本文的数据集来源于10个纯合亚种的栽培稻,共计1 414个样本,每个亚种的样本数如表1所示,每个样本位点数为3 139 562对碱基对,经过预处理后,每个样本位点数为3 139 562个碱基。样本使用标签编码进行编码,在1.1.2节中已经描述了编码方式,将样本的类别标签使用0~9进行编码。从每个品种中随机选择10%的数据集作为测试集,用于测试模型的分类准确率,剩余的90%作为训练集用于训练模型。

2.2 利用改进的卡方选择SNP子集

根据1.2.1节描述的算法计算样本中每一个位点的卡方值,然后选择卡方值最高的前1%的位点,标记它的索引。经过卡方计算后,SNP位点从3 139 562个减少到31 396个。然后使用R语言的Cmplot包查看SNP位点降维前后在12条染色体上的分布密度,两者的分布情况如图4所示。从图4两个子图的对比分析中可以看到,降维前3 139 562个SNP位点均匀分布在12条染色体上,而降维后,SNP位点主要集中在1号、6号和12号染色体上。这说明原始样本中存在着大量冗余的SNP位点,不利于分类器分类样本,通过改进的卡方选择的SNP子集包含了与分类高相关的信息量,减少了分类器的计算量。

本文使用最小等位基因频率(Minor Allele Frequency,MAF)评估改进的卡方模型的有效性。MAF通常是指在给定种群中的不常见的等位基因发生频率,如果MAF太小甚至等于0,说明该基因位点上的基因没有分型,不利于分类,应该删除。一般来说MAF的阈值为0.05或0.01,这是根据样本的大小决定的。本文使用R语言查看选择的SNP子集的MAF分布,如图5所示,本文的模型过滤掉了99% MAF<0.01的基因位点,SNP子集中有93.6%的基因位点的MAF>0.05,说明所述的过滤方法标记基因分型非常有效。

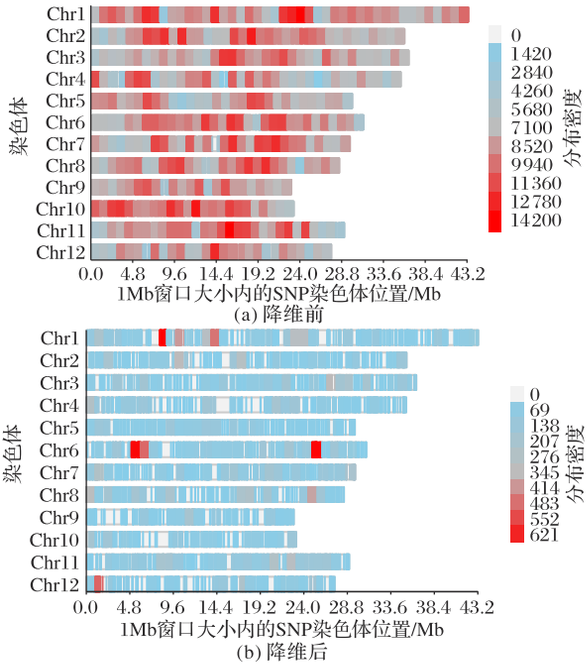


图4 SNP位点降维前后在12条染色体上的分布

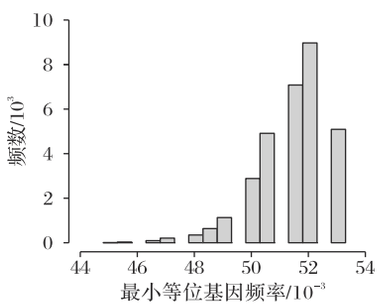


图5 31 396个SNP位点的MAF分布

2.3 Variant-SMOTE平衡样本及分类结果分析

本文使用留出法按9:1将SNP子集划分为两个互斥的数据集,其中90%的数据集作为训练集,10%的数据集作为测试集,即测试集142个样本,训练集1 272个样本。划分后的数据集分布与原始数据集分布保持一致性,然后对训练集进行样本平衡,平衡后训练集样本量为3 610,测试集不做改变,保证使用真实测试集验证本文方法。

10个亚种在测试集和训练集中的分布如表5所示。

表5 10个亚种在测试集和训练集中的分布

亚种名称	平衡前样本数		平衡后样本数	
	训练集	测试集	训练集	测试集
Aromatic	66	3	361	3
Aus	39	3	361	3
Indica1	51	5	361	5
Indica2	26	5	361	5
Indica3	137	14	361	14
IndicaIntermediate	81	8	361	8
Intermediate	31	5	361	5
JaponicaIntermediate	145	20	361	20
TemperateJaponica	335	40	361	40
TropicalJaponica	361	39	361	39

由于平衡样本之前,测试集中TemperateJaponica的样本量与基准量只差1个,所以本文设置k=1,利用Variant-SMOTE算法将SNP子集进行样本平衡,使用KNN查找每个少数类样本的1个近邻,并生成一个合成样本。然后将Aromatic、Aus、Indica1、Indica2、Indica3、IndicaIntermediate、Intermediate、JaponicaIntermediate、TemperateJaponica这9个类别样本量少于基准值的亚种增加至基

中国知网 https://www.cnki.net

准值。最后,将平衡后的样本输入k=3的KNN分类器中进行训练。

本文使用ReliefF和chi-square两种特征选择方法从水稻SNP数据集中分别选出31 396个重要SNP位点,同样地,使用留出法将数据集分成90%的训练集和10%的测试集,然后在三种分类器上训练10次,分类准确率如图6所示。

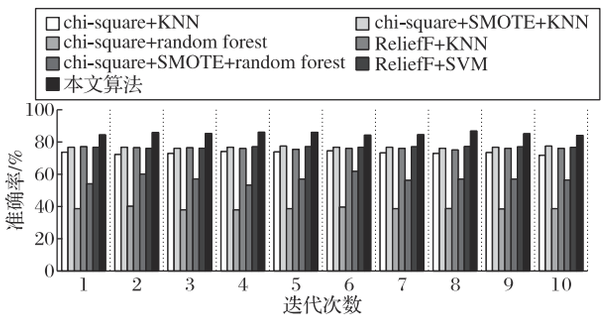


图6 本文方法与其他方法的分类准确率对比

6种对比方法[18-23]中,ReliefF和SVM融合效果最好,准确率最高可以达到77.05%,最低为75.01%;chi-square和随机森林(random forest)的融合效果最差,最高的准确率只有40.14%;而本文方法的准确率最高可达86.75%,最低为84.25%。

从表6所示的平均值对比结果中可看出,使用卡方筛选重要SNP位点的方法分类效果最差,然而,chi-square在KNN分类器上表现出更好的效果,说明卡方更适合与KNN相结合。相较于文献[18-23]方法,本文方法在水稻数据集上的鉴定效果最好,平均分类准确率为85.21%,比最低的方法[19]提高了120%,比最高的方法[22-23]提高了11.01%。本文还比较了时间性能,本文方法牺牲了时间性能,该部分时间性能消耗主要是在数据平衡上。

表6 不同方法在水稻数据集上的分类性能比较(平均值)

方法	准确率/%	耗时/min
chi-square+SMOTE+random forest[18]	57.00	107.78
chi-square+random forest[19]	38.73	83.76
ReliefF+KNN[20]	76.05	45.51
chi-square+KNN[21]	73.24	84.12
chi-square+SMOTE+KNN[22]	76.76	99.89
ReliefF+SVM[23]	76.76	36.95
本文方法	**85.21**	**109.98**

3 结语

本文方法利用了改进的卡方、Variant-SMOTE和KNN构建了一个用于鉴定水稻亚种的分类器,该模型可以从超高维度的数据集中筛选与亚种分类相关性最高的SNP位点子集,以区分栽培稻亚种。实验结果表明,本文方法对栽培稻的亚种鉴定准确率达到了85.21%。相较于其他的降维方法,本文方法在栽培稻数据集中对SNP位点的筛选更有效,亚种鉴定效果优于对比方法。虽然本文方法在纯合子的SNP位点筛选上很有效,但是却在时间性能表现上不占优势,而且对于其他物种和杂合子的品种鉴定是否也有这么好的效果也还需要进一步实验验证。

参考文献 (References)

[1] 刘月丽,覃锡忠,贺三刚,等. PCA与随机森林相结合筛选高信息量SNP位点:应用于羊的品种鉴定[J]. 计算机工程与应用,2018,54(16):235-240.

[2] 郑向华,叶俊华,程朝平,等. 利用SNP标记进行水稻品种籼粳鉴定[J]. 作物学报, 2022, 48(2):342-352.

[3] YUAN X, LI Z, XIONG L, et al. Effective identification of varieties by nucleotide polymorphisms and its application for essentially derived variety identification in rice[J]. BMC Bioinformatics, 2022, 23(1): 1-16.

[4] 戚华雄,查中萍. 水稻品种籼粳属性的鉴定方法研究[J]. 湖北农业科学,2015,54(24):6130-6133.

[5] DONG W, CHEN J W, SHEN L Y, et al. The development of a vision-based phenotypic analysis algorithm for measuring spikelet-related traits in rice[J]. Plant Physiology Journal, 2022, 58(5): 957-971.

[6] 王彩红,徐群,于萍,等. 亚洲栽培稻程氏指数与SSR标记分类的比较分析[J]. 中国水稻科学,2012,26(2):165-172.

[7] WEN W W, MEI H W, FENG F J, et al. Population structure and association mapping on chromosome 7 using a diverse panel of Chinese germplasm of rice (Oryza sativa L.)[J]. Theoretical and Applied Genetics, 2009, 119(3): 459-470.

[8] LI L, CAO L M. SNP selection based on k-center algorithm [C]// Proceedings of the 2019 6th International Conference on Systems and Informatics. Piscataway: IEEE, 2019: 1449-1454.

[9] MEI H W, FENG F J, LU B R, et al. Experimental validation of inter-subspecific genetic diversity in rice represented by the differences between the DNA sequences of 'Nipponbare' and '93-11'[J]. Chinese Science Bulletin, 2007, 52(10): 1327-1337.

[10] ALZUBI R, RAMZAN N, ALZOUBI H, et al. A hybrid feature selection method for complex diseases SNPs [J]. IEEE Access, 2017, 6: 1292-1301.

[11] EVANS D T. A SNP microarray analysis pipeline using machine learning techniques[D]. Athens: Ohio University, 2010: 34.

[12] BATNYAM N, GANTULGA A, OH S. An efficient classification for Single Nucleotide Polymorphism (SNP) dataset [M]// LEE R. Computer and Information Science. Heidelberg: Springer, 2013: 171-185.

[13] GUNASEKARAN H, RAMALAKSHMI K, REX MACEDO AROKIARAJ A, et al. Analysis of DNA sequence classification using CNN and hybrid models[J]. Computational and Mathematical Methods in Medicine, 2021, 2021: 1835056.

[14] KARAGÖZ M A, NALBANTOGLU O U. Taxonomic classification of metagenomic sequences from relative abundance index profiles using deep learning[J]. Biomedical Signal Processing and Control, 2021, 67: 102539.

[15] ZHAI Y J, SONG W, LIU X J, et al. A chi-square statistics based feature selection method in text classification [C]// Proceedings of the 2018 IEEE 9th International Conference on Software Engineering and Service Science. Piscataway: IEEE, 2018: 160-163.

[16] FERNÁNDEZ A, GARCIA S, HERRERA F, et al. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary[J]. Journal of Artificial Intelligence Research, 2018, 61: 863-905.

[17] ZHANG S, ZHU F, YU Q, et al. Identifying DNA-binding proteins based on multi-features and LASSO feature selection[J]. Biopolymers, 2021, 112(2): e23419.

[18] SHARMA K, CHAWLA R, SHUKLA U, et al. Predictive analysis based on feature relevance estimation for survival rate post heart failure using ensemble learners [C]// Proceedings of the 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics. Piscataway: IEEE, 2022: 1-7.

[19] ONESIME M, YANG Z Y, DAI Q. Genomic island prediction via chi-square test and random forest algorithm [J]. Computational and Mathematical Methods in Medicine, 2021, 2021: 9969751.

[20] ALOTAIBI A S. Hybrid model based on ReliefF algorithm and K-nearest neighbor for erythemato-squamous diseases forecasting [J]. Arabian Journal for Science and Engineering, 2022, 47(2): 1299-1307.

[21] MUSHTAQ Z, YAQUB A, SANI S, et al. Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets[J]. Journal of the Chinese Institute of Engineers, 2020, 43(1): 80-92.

[22] SREEJITH S, NEHEMIAH H K, KANNAN A. Clinical data classification using an enhanced SMOTE and chaotic evolutionary feature selection [J]. Computers in Biology and Medicine, 2020, 126: 103991.

[23] SAHU S, DAS P, BINU JOSE A. A new framework for disease prediction: using dimensionality reduction and feature selection [M]// KUMAR R, VERMA A K, SHARMA T K, et al. Soft Computing: Theories and Applications. Singapore: Springer Nature Singapore, 2023: 381-396.

中国知网 https://www.cnki.net

Computer Science
中国计算机学会会刊
ISSN 1002-137X
CN 50-1075/TP
CODEN JKIEBK

计算机科学

2022年 第49卷 第6A期
Vol.49 No.6A 2022



本期封面：舆论动力学模型综述
刘淑娴，徐欢，王微，邓乐

计算机科学

（JISUANJI　KEXUE）

第 49 卷　第 6A 期　2022 年 6 月

目　次

智慧医疗

1　指静脉识别技术研究综述　刘伟业　鲁慧民　李玉鹏 等

12　基于多尺度特征的脑肿瘤分割算法　孙福权　崔志清　邹　彭 等

17　基于 Transformer 和 LSTM 的药物相互作用预测　康　雁　徐玉龙　寇勇奇 等

22　基于深度学习的黑色素瘤智能诊断多模型算法　常炳国　石华龙　常雨馨

27　蓝舌病毒基因组序列多元概率特征可视化分析　陈慧嫔　王　琨　杨　恒 等

32　一种新的中文电子病历文本检索模型　于家畦　康晓东　白程程 等

39　医疗 CPS 协作网络控制策略优化　刘　丽　李仁发

44　基于胸腔积液超声图像标准化方法的胸腔积液性质分析模型　冯亦凡　徐　琪　曾卫明

54　基于改进 CNN 的低剂量 CT 图像肺结节自动检测　岳　晴　尹健宇　王生生

60　基于注意力机制和多任务学习的阿尔茨海默症分类　杜丽君　唐玺璐　周　娇 等

智能计算

66　自然语言处理在简历分析中的应用研究综述　李小伟　舒　辉　光　焱 等

74　知识图谱推理研究综述　马瑞新　李泽阳　陈志奎 等

86　面向超参数估计的贝叶斯优化方法综述　李亚茹　张宇来　王佳晨

93　多示例学习算法综述　赵　璐　袁立明　郝　琨

100　一种可快速迁移的领域知识图谱构建方法　邓　凯　杨　频　李益洲 等

109　命题逻辑中三元子句集的冗余文字　李　洁　钟小梅

113　一种基于支持向量机的主动度量学习算法　侯夏晔　陈海燕　张　兵 等

119　基于跨句上下文信息的神经网络关系分类方法　黄少滨　孙雪薇　李熔盛

125　混合改进的花授粉算法与灰狼算法用于特征选择　康　雁　王海宁　陶　柳 等

133　军事指控保障领域命名实体识别语料库的构建　杜晓明　袁清波　杨　帆 等

140　基于剩余格的模糊粗糙集的拓扑性质　许思雨　秦克云

144　基于不平衡数据与集成学习的属性级情感分类　林　夕　陈孜卓　王中卿

150　融合 Bert 和图卷积的深度集成学习软件需求分类　康　雁　吴志伟　寇勇奇 等

159　利用粒子滤波方法求解数据包络分析问题　黄国兴　杨泽铭　卢为党 等

165　一种融合多层次情感和主题信息的 TS-AC-EWM 在线商品排序方法　余本功　张子薇　王惠灵

172　基于图卷积网络的专利摘要自动生成研究　李健智　王红玲　王中卿

178 基于动量的映射式梯度下降算法 吴子斌 闫 巧
184 基于约束优化生成式对抗网络的数据去偏方法 徐国宁 陈奕芃 陈一鸣 等
191 考虑一单多品的外卖订单配送时间的带时间窗的车辆路径问题 杨浩雄 高 晶 邵恩露
199 基于四边形最优圈内最短路径的旅行商问题割边方法 王 永 崔 源
206 TI-FastText 自动商品分类算法 邵欣欣
211 基于改进麻雀搜索优化支持向量机的渔船捕捞方式识别 单晓英 任迎春
217 基于多种改进策略的改进麻雀搜索算法 李丹丹 吴宇翔 朱聪聪 等
223 具有多个突触通道的新型数值脉冲神经 P 系统的计算能力研究 殷 秀 刘希林 刘希玉

大数据&数据科学

232 基于 DBSCAN 聚类的集群联邦学习方法 鲁晨阳 邓 苏 马武彬 等
238 一种改进的融合相似度和信任度的协同过滤算法 蔡晓娟 谭文安
242 自适应的集成定序算法 王文强 贾星星 李 朋
247 基于隐马尔可夫模型的铁路出行团体关系预测研究 王 欣 向明月 李思颖 等
256 SDFA:基于多特征融合的船舶轨迹聚类方法研究 郁舒昊 周 辉 叶春杨 等
261 基于 DE-LSTM 模型的教育统计数据预测研究 刘宝宝 杨菁菁 陶 露 等
267 基于用户场景的 Android 应用服务推荐方法 王 毅 李政浩 陈 星
272 基于点割集图分割的矩阵变换与分解的推荐算法 何亦琛 毛宜军 谢贤芬 等
280 基于网络媒体的非线性动力学信息传播模型 杜鸿毅 杨 华 刘艳红 等
285 基于密度敏感距离和模糊划分的改进 FCM 算法 毛森林 夏 镇 耿新宇 等
291 基于隐半马尔可夫模型的微博流行信息检测方法 谢柏林 黎 琦 邝 建
297 一种适于多分类问题的支持向量机加速方法 陈景年

图像处理&多媒体技术

301 语音风格迁移研究进展 刘 畅 魏为民 孟繁星 等
309 基于 XGBoost 算法的水稻基因组 6mA 位点识别研究 孙福权 梁 莹
314 视频流的实时景深延拓算法 来腾飞 周海洋 余飞鸿
319 基于多源迁移学习的大坝裂缝检测 王君锋 刘 凡 杨 赛 等
325 基于极大极稳定区域及 SVM 的交通标志检测 胡 聪 何晓晖 邵发明 等
331 智能语音技术端到端框架模型分析和趋势研究 李 荪 曹 峰
337 基于 YOLOv4 的目标检测知识蒸馏算法研究 楚玉春 龚 航 王学芳 等
345 融合交叉注意力机制的图像任意风格迁移 杨 玥 冯 涛 梁 虹 等
353 一种结合双注意力机制和层次网络结构的细碎农作物分类方法 杨健楠 张 帆
358 基于改进 YOLOv3 的空间非合作目标部件识别算法 郝 强 李 杰 张 曼 等
363 改进注意力机制的多叉树网络多作物早期病害识别方法 高荣华 白 强 王 荣 等
370 一种基于 Bottleneck Transformer 的轻量级微表情识别架构 张嘉淏 刘 峰 齐佳音
378 改进 Faster R-CNN 的光学遥感飞机目标检测 祝文韬 兰先超 罗唤霖 等
384 基于虚拟曝光方法的单幅逆光图像增强 赵明华 周童童 都双丽 等

390 基于纹理特征增强和轻量级网络的人脸防伪算法 沈 超 何希平
397 基于多源数据和逻辑推理的行为识别技术研究 肖治鸿 韩晔彤 邹永攀
407 基于局部正则二次线性重构表示的人脸识别 黄 璞 杜旭然 沈阳阳 等
412 基于 GDIoU 损失函数的 YOLOv4 绝缘子高效定位算法 马 宾 付永康 王春鹏 等
418 不同数据增强方法对模型识别精度的影响 王建明 陈响育 杨自忠 等
424 基于外接圆半径差损失的实时安全帽检测算法 陈永平 朱建清 谢 懿 等
429 基于局部约束特征线表示的人脸识别 黄 璞 沈阳阳 杜旭然 等
434 基于离散小波变换的双域特征融合深度卷积神经网络 孙洁琪 李亚峰 张文博 等
441 基于 CUDA 核函数的多路视频图像拼接加速算法 刘 云 董守杰

信息安全

447 区块链技术的研究及其发展综述 傅丽玉 陆歌皓 吴义明 等
462 RegLang:一种面向监管的智能合约编程语言 高健博 张家硕 李青山 等
469 基于智能合约的秘密重建协议 卫宏儒 李思月 郭涌浩
474 Python 虚拟机本地代码的安全性实证研究 蒋成满 华保健 樊淇梁 等
480 面向网络安全训练评估的受训者行为描述模型 陶礼靖 邱 菡 朱俊虎 等
485 抵御计时攻击的基于 Knuth-Yao 的二元离散高斯采样算法 梁懿雯 杜育松
490 基于 Fabric 的电子病历跨链可信共享系统设计与实现 袁昊男 王瑞锦 郑博文 等
496 一种提高联邦学习模型鲁棒性的训练方法 闫 萌 林 英 聂志深 等
502 比特币实体交易模式分析 何 茜 贺可太 王金山 等
508 一种基于异质模型融合的 Android 终端恶意软件检测方法 姚 烨 朱怡安 钱 亮 等
516 基于双向蚁群算法的网络攻击路径发现方法 高文龙 周天阳 朱俊虎 等
523 符合监管合规性的自动合成新闻检测方法研究 毛典辉 黄晖煜 赵 爽
531 具有仿冒攻击检测的鲁棒性说话人识别 郭星辰 俞一彪
537 基于医疗联盟链的跨域认证方案设计 陈彦冰 钟超然 周超然 等
544 基于半监督学习的网络流量分析研究 庞兴龙 朱国胜
555 基于 Renyi 熵和 BiGRU 算法实现 SDN 环境下的 DDoS 攻击检测方法 杨亚红 王海瑞
562 基于 SMOTE-SDSAE-SVM 的车载 CAN 总线入侵检测算法 周志豪 陈 磊 伍 翔 等
571 一种基于热带半环的密钥建立协议的安全性分析 黄华伟 李春华
575 基于隐私保护的反向传播神经网络学习算法 王 健
581 未知网络攻击识别关键技术研究 曹扬晨 朱国胜 孙文和 等
588 基于进化神经网络的电力信息网安全态势量化方法 吕鹏鹏 王少影 周文芳 等

计算机网络

594 面向无尺度图的 Δ-stepping 算法改进策略 陈钧吾 余华山
601 D2D 辅助移动边缘计算下的卸载策略优化 方 韬 杨 旸 陈佳馨
606 基于无迹粒子滤波的 WiFi-PDR 融合室内定位技术 周楚霖 陈敬东 黄 凡
612 一种基于有向感知区域调整的强栅栏构建算法 王方红 范兴刚 杨静静 等

619 多无人机使能移动边缘计算系统中的计算卸载与部署优化 刘漳辉 郑鸿强 张建山 等
628 改进灰狼算法的无线传感器网络覆盖优化 范星泽 禹 梅
632 存在 CSI 估计错误的增强型 ELM 叠加 CSI 反馈方法 卿朝进 杜艳红 叶 青 等
639 基于服务器集群的负载均衡优化调度算法 田真真 蒋 维 郑炳旭 等

交叉&应用

645 早期量子算法在量子通信、量子纠错等领域的应用 Renata WONG
649 Shor 整数分解算法的线路优化 刘建美 王 洪 马 智
654 面向供应链风险评估的改进 BP 小波神经网络研究 徐佳楠 张天瑞 赵伟博 等
661 机械零件加工工艺本体库构建 王钰珏 梁宇豪 王素琴 等
667 基于螺旋进化萤火虫算法和 BP 神经网络的模型及其在 PPP 融资风险预测中的应用 朱旭辉 沈国娇 夏平凡 等
675 基于 CNN-LSTM 的卫星云图云分类方法研究 王 杉 徐楚怡 师春香 等
680 数据科学与大数据技术课程体系的复杂网络分析 杨 波 李远彪
686 基于改进人工势场的未知障碍物无人机编队避障 陈博琛 唐文兵 黄鸿云 等
694 基于前景理论的微信健康信息质量三方博弈分析 王显芳 张 亮 张 宁
705 支持多租户模式的业务流程动态定制模型 张纪林 邵玉曹 任永坚 等
714 基于加速度模糊控制的六足机器人遥操作 尹宏俊 邓 楠 程亚迪
723 面向食品溯源场景的 PBFT 优化算法应用研究 李 博 向海昀 张宇翔 等
729 基于改进准深度算法的诊断策略优化方法 张志龙 史贤俊 秦玉峰
733 融合用户和区位资源特征的混合房源推荐方法 朴 勇 朱锶源 李 阳
738 物联网僵尸网络病毒的传播动力学模型与分析 张翕然 刘万平 龙 华
744 考虑信息泄露和信息投资成本因素下平台的信息共享及渠道策略研究 徐明月
753 基于模糊双目标规划的充电站布局模型 阙华坤 冯小峰 郭文翀 等
759 基于步态分类辅助的虚拟 IMU 的行人导航方法 杨 涵 万 游 蔡洁萱 等
764 基于法线迭代的模型中轴生成方法 宗迪迪 谢益武
771 适用于各单元共识交易的电力区块链系统优化调度研究 周 航 姜 河 赵 琰 等
777 向量 DSP 的混合资源启发式循环展开因子选择方法研究 陆浩松 胡勇华 王书盈 等
784 基于 Stacking 多模型融合的 IGBT 器件寿命的机器学习预测算法研究 王 飞 黄 涛 杨 晔
790 Grassberger 熵随机森林在窃电行为检测的应用 阙华坤 冯小峰 刘盼龙 等
795 海上风电场通用运维路径规划模型优化及仿真 谭任深 徐龙博 周 冰 等
802 基于 CAN 总线纯电动汽车的整车控制器 陈龙华 李洪林 杨汉立 等

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210600159

基于点割集图分割的矩阵变换与分解的推荐算法

何亦琛[1]　毛宜军[1]　谢贤芬[2]　古万荣[1]
1 华南农业大学数学与信息学院　广州 510642
2 暨南大学经济学院　广州 510632
(heyichen666@hotmail.com)

摘　要　基于模型的协同过滤算法通过矩阵分解来将用户偏好以及物品属性用隐变量来表示,但现有的矩阵分解算法很难应对个性化推荐系统中严重的数据稀疏性以及数据变化性所带来的问题。针对上述问题,提出了基于双边块对角矩阵的矩阵分解算法。首先通过基于社区发现的点割集图分割算法将原始的稀疏矩阵转换成双边块对角矩阵,将具有相同偏好的用户以及相似特征的物品归并到同一个对角块中,然后将结构中的对角块和双边拼接成数个密度较高的子矩阵。实验结果表明,对这几个密度有提高的子矩阵进行并行分解,基于其分解结果进行原始矩阵的预测,能够有效缓解矩阵分解中数据稀疏性所带来的问题,既能提升预测的精度,又能提高推荐结果的可解释性。同时,每个子对角块都能并行独立分解,能达到提高算法效率的目的。

关键词:社区发现;矩阵分解;协同过滤;推荐系统
中图法分类号　TP391.3

Matrix Transformation and Factorization Based on Graph Partitioning by Vertex Separator for Recommendation

HE Yi-chen[1],MAO Yi-jun[1],XIE Xian-fen[2] and GU Wan-rong[1]
1 School of Mathematics and Information,South China Agricultural University,Guangzhou 510642,China
2 School of Economy,Jinan University,Guangzhou 510632,China

Abstract　Model-based collaborative filtering algorithms usually express user's preferences and item's attributes by latent factors through matrix factorization,but the traditional matrix factorization algorithm is difficult to deal with the serious data sparsity and data variability problems in the recommendation system. To solve the above problem,a matrix factorization algorithm based on bordered block diagonal matrix is proposed. Firstly,the original sparse matrix is transformed into bordered block diagonal matrix by a graph partitioning algorithm based on community discovery,which merges users with the same preference and items with similar characteristics into the same diagonal block,and then splices the diagonal blocks and the bordered into several sub-diagonal matrices which have higher densities. The experimental results show that,by decomposing the sub-diagonal matrices in parallel can not only improve the precision of prediction,but also improve the interpretability of the recommendation results. At the same time,each sub-diagonal matrix can be decomposed independently and in parallel,which can improve the efficiency of the algorithm.

Keywords　Community discovery,Matrix factorization,Collaborative filtering,Recommendation system

1　引言

随着互联网的高速发展,人们在线上的活动也变得越来越频繁,因此产生了庞大且复杂的数据集。如何利用这些庞大的数据集将个性化推荐应用到不同的场景是当下需要解决的热点问题。

协同过滤算法充分利用集体的智慧,通过分析大量用户的行为来挖掘某些隐含的模式,基于协同过滤的推荐算法得到了最广泛以及最深入的研究[1-2]。基于协同过滤的推荐系统又可以细分为基于用户的协同过滤系统、基于物品的协同过滤系统、基于模型的协同过滤系统[3-5]。其中,基于隐变量模型的协同过滤是最为高效的协同过滤算法之一[6],基于矩阵分解的协同过滤算法将用户偏好用少量的隐变量来表示,矩阵分解算法的思想都是将原始矩阵分解为子矩阵,利用子矩阵的乘积对原始矩阵进行还原及预测。常见的矩阵分解算法包括 SVD(奇异值分解)[7]、NMF(非负矩阵分解)[8]、

基金项目:全国统计科学研究项目(2020LY018);全国统计科学研究重点项目(2019LZ37);广东省社会科学项目(GD19CGL34);国家重点研发计划(2017YFC1601701);广东省农业厅创新团队项目(2019KJ130)
This work was supported by the National Statistical Science Research Project(2020LY018),National Statistical Science Research Key Project(2019LZ37),Social Science Project of Guangdong Province(GD19CGL34),National Key Research and Development Project(2017YFC1601701) and Department of Agriculture of Guangdong Provincial Innovation Research Team Project(2019KJ130).
通信作者:古万荣(guwanrong@scau.edu.cn)

PMF(概率矩阵分解)[9]等等。

目前的矩阵分解算法虽然在推荐系统应用的预测精度方面有着较好的表现，但是依旧面临着以下问题：

(1)数据的稀疏性，在实际的系统中，通常用户只对系统中极个别的物品进行过评分，若直接对整个原始矩阵进行矩阵分解，则必然会带来计算量庞大的问题。

(2)数据的变化性，在实际的系统中，用户的行为数据会不断变化，模型的局部变化必然会对整个模型带来改变，如何对模型进行重复训练是需要解决的问题。

(3)推荐结果的可解释性，矩阵分解的分解结果是基于隐变量的，但并不能很明确地确定这些隐变量的具体含义，很难给予用户一个可接受且信服的推荐结果。

为解决上述问题，本文提出了基于双边块对角矩阵的矩阵分解算法，主要贡献可以总结如下：

(1)受社区发现算法启发，本文利用基于社区发现的点割集图分割算法，对原始稀疏矩阵进行划分，得到双边块对角矩阵结构，该结构将具有相似偏好的用户和具有相似属性的物品聚集在一起，提升原始矩阵的密度，缓解数据稀疏性带来的问题，基于该结构进行矩阵分解并进行协同过滤，也可以对推荐结果进行更好的解释。

(2)本文提出了基于上述结构的近似矩阵分解算法，实验结果表明，相比 4 种基线算法，本文提出的基于双边块对角矩阵的矩阵分解算法的预测精度以及速度都有较大幅度的提高。同时，当原始矩阵的某处评分发生变化后，不再需要对整个模型进行重新分解，只需要在发生变换的子矩阵上重新分解即可，从而大大提升了在实际推荐系统的分解效率以及可扩展性。

2 相关工作

为了解决数据稀疏性的问题，目前较为流行的方法为矩阵聚类算法[10-12]，该算法预先将用户的行向量以及物品的列向量进行聚类，然后根据聚类的结果，将预测限定在特定的类别之中，以达到降低预测计算量的目的。但是聚类算法也有以下不足之处，通常来说，聚类的结果难以给出令用户信服的解释，并且聚类算法将系统中的一个用户或一个物品强制归属于一个类中，这与实际的系统也是相违背的。另一种数据降维方法 PCA 方法(主成分分析)是以 SVD 为基础发展起来的方法，PCA 方法的思想是把用户和物品映射到另外一个潜在特征空间中，在这个空间中，用户和物品的最主要特征可以很轻易地表达出来。已有学者，如 Aaldering[13]直接将 PCA 应用到协同过滤之中。但是，PCA 方法也有如下的缺点：PCA 方法无法通过参数化等方法对处理过程进行干预，假设系统开发人员已对需要进行预测的内容有一定的先验知识，但是无法调节其方法，这样可能导致预测效果不理想，同时算法效率也会受到影响。

为了解决模型需要重复训练的问题，学者们对流行的矩阵分解算法进行了拓展研究，如 Nguyen[14]提出了一种分布式机器学习方法(DE-SVD)，该算法使用边缘服务器进行数据初步处理，数据的最终处理则由集中式服务器完成。还有 Yu 等[15]提出使用坐标下降法代替交替二乘法(ALS)和随机梯度下降法(SGD)去计算矩阵分解问题的最优解。Liu 等[16]提出了基于 SVD 填充的混合推荐算法，也通过随机梯度法去填充稀疏矩阵，并且将时间权重融入矩阵中以优化用户相似度。他们的目的都是为了解决在频繁交互下模型可扩展性不足的问题。但是这些算法在矩阵的局部发生变化时，仍需要对整个模型进行分解，在数据量庞大的情况下，其计算量依旧非常大。

为了更好地对用户和物品进行分类，并给出令用户信服的推荐结果，社区发现算法开始得到了学者们的研究[17-20]。聚类算法和社区发现算法具有相似之处，但是两种算法处理的对象不同，聚类算法所处理的是能表示为高维向量的属性数据，而社区发现算法所处理的是能表示为网络的关系数据。近来，也有不少学者将社区发现算法与矩阵分解算法结合起来进行研究，如 Gligorijević 等[21]将 NMF 算法和社区发现结合起来进行研究，提高了矩阵分解结果的可解释性。

3 基于社区发现的点割集图分割算法

对于无向图 $G(V,E)$，有顶点集 $V'\subset V$，$V''\subset V'$，如果删除 V'，则 G 不再连通，但如果删除 V' 的子集 V'' 后图依旧是连通图，则称 V' 是无向图 G 的一个点割集。

每一个稀疏的矩阵都可以用一个二部图来表示(见图 1)，假设对原始评分矩阵对应的二部图执行基于点割集的图分割算法，通过移除所有的点割集，以及点割集顶点对应的边，便可以得到 k 个互不连通的连通分支，如图 2 所示。

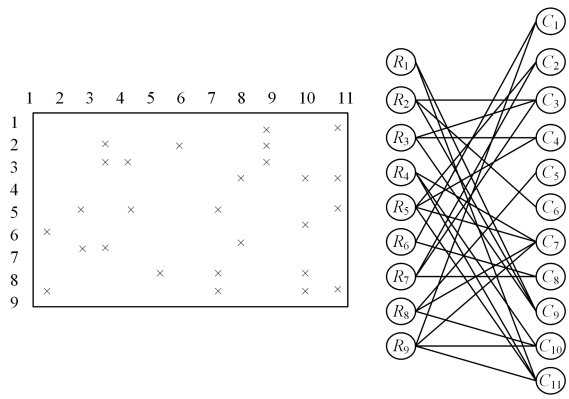


图 1 原始矩阵及其二部图

Fig. 1 Original matrix and its bipartite graph

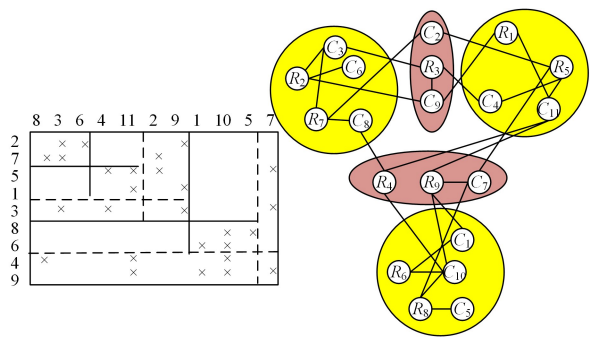


图 2 双边块对角矩阵及其二部图(电子版为彩图)

Fig. 2 Bordered block diagonal matrix and its bipartite graph

图 2 中，黄色部分为无向图的 3 个连通分支，红色部分为无向图的点割集。这 k 个连通分支对应到本文提出的双边块对角矩阵上，即 k 个对角块，点割集对应到双边块对角矩阵上的双边上。双边块对角结构可以抽象表示为：

$$X=\begin{array}{c}\begin{matrix}J_1 & J_2 & J_3\end{matrix}\\ \begin{bmatrix}D_1 & & C_1\\ & D_2 & C_2\\ R_1 & R_2 & B\end{bmatrix}\end{array}\begin{matrix}I_1\\ I_2\\ I_3\end{matrix}=\begin{array}{c}\begin{matrix}J_{11} & J_{12} & J_{B1} & J_2 & J_B\end{matrix}\\ \left[\begin{array}{ccc:c:c}D_{11} & & & & C_{11}^1\\ & D_{12} & & & C_{12}^2\\ R_{11} & R_{12} & B_1 & & C_1^3\\ \hdashline & & & D_2 & C_2\\ \hdashline R_1^1 & R_1^2 & R_1^3 & R_2 & B\end{array}\right]\end{array}\begin{matrix}I_{11}\\ I_{12}\\ I_{B1}\\ I_2\\ I_B\end{matrix} \quad (1)$$

其中，D 称为对角块，R 称为行边，C 称为列边，B 为双边交叉块，因此这个矩阵被称为双边块对角矩阵。

点割集的意义在于，能构成点割集的都是“超级用户”或者“超级物品”，对应到双边 R，C 上，这些用户和物品跨社区行为十分频繁，直观地来说，就是被排列到双边的用户兴趣十分广泛，无论什么类型的物品他都会有所涉猎，超级物品在实际的系统中即受到大多数用户喜爱的热门物品。不同的连通分支表示具有相同偏好的用户以及相似属性的物品所构成的集合，通过该算法得到的对角块 D，就是同一类用户和物品的集合，这就是基于二部图的社区发现图分割算法的意义所在。

进一步地，将双边块对角矩阵的双边排列与对角块排列在一起，即可得到数个子矩阵，例如将式(1)中的双边块对角矩阵排列为块对角矩阵，可以得到：

$$X=\begin{array}{c}\begin{matrix}J_{11} & J_{B1} & J_B & J_{12} & J_{B1} & J_B & J_2 & J_B\end{matrix}\\ \left[\begin{array}{ccc:ccc:c:c}D_{11} & C_{11} & C_1^1 & & & & & \\ R_{11} & B_1 & C_1^3 & & \tilde{X}_{12} & & \tilde{X}_{13} & \\ R_1^1 & R_1^3 & B & & & & & \\ \hdashline & & & D_{12} & C_{12} & C_1^2 & & \\ \tilde{X}_{21} & & & R_{12} & B_1 & C_1^3 & \tilde{X}_{23} & \\ & & & R_1^2 & R_1^3 & B & & \\ \hdashline & & & & & & D_2 \ C_2 & \\ \tilde{X}_{31} & & & \tilde{X}_{32} & & & R_2 \ B & \\ & & & & & & & \end{array}\right]\end{array}\begin{matrix}I_{11}\\ I_{B1}\\ I_B\\ I_{12}\\ I_{B1}\\ I_B\\ I_2\\ I_B\end{matrix}$$

$$\triangleq \mathrm{diag}(\tilde{X}_1, \tilde{X}_2, \tilde{X}_3) \quad (2)$$

其中，当 $i\neq j$ 时，$X_{ij}=0$。本文基于这些对角线上的子对角块矩阵进行独立的矩阵分解，通过对重复预测的矩阵求平均，即可近似地去恢复原始矩阵。

4 基于双边块对角矩阵的矩阵分解

原始矩阵通过执行基于社区发现的点割集图分割算法，可以得到双边块对角矩阵，其中双边块对角矩阵中的每个对角块都是具有相同偏好的用户以及具有相似属性的物品的集合，如何利用这个重要的信息去进行矩阵分解，以提升预测的精度是本文需要解决的问题。

本节中，需要解决的问题如下：

(1)如何保证块对角矩阵中的子矩阵的规模是相似的。

(2)如何对每个规模相似的子矩阵进行矩阵分解。

(3)如何利用子矩阵的分解结果去预测块对角矩阵中的零块，以及恢复原始矩阵。

下文将详细阐述基于双边块对角矩阵精确恢复原始矩阵的思想，对块对角矩阵中的每个子对角块进行分解，并利用分解结果去近似原始矩阵。

4.1 基于双边块对角矩阵精确恢复原始矩阵

原始矩阵虽然通常不能直接转换为块对角矩阵，但是几乎都可以被转换为双边块对角矩阵，假设原始矩阵 X 被转换为双边块对角矩阵 $\tilde{X}$，其中 $\tilde{X}$ 可以被分解为：

$$\tilde{X}\triangleq\begin{bmatrix}D_i & C_i\\ R_i & B\end{bmatrix}=U_iV_i^{\mathrm{T}}=\begin{bmatrix}U_{i1}\\ U_{i2}\end{bmatrix}\begin{bmatrix}V_{i1}^{\mathrm{T}} & V_{i2}^{\mathrm{T}}\end{bmatrix} \quad (3)$$

根据式(3)，易得 $D_i=U_{i1}V_{i1}^{\mathrm{T}}$，$R_i=U_{i2}V_{i1}^{\mathrm{T}}$，$C_i=U_{i1}V_{i2}^{\mathrm{T}}$，$B=U_{i2}V_{i2}^{\mathrm{T}}$，令：

$$U=\begin{bmatrix}U_{11} & & & \\ & U_{21} & & \\ & & \ddots & \\ & & & U_{k1}\\ U_{12} & U_{22} & \cdots & U_{k2}\end{bmatrix},=\begin{bmatrix}V_{11} & & & \\ & V_{21} & & \\ & & & \\ & & \ddots & V_{k1}\\ V_{12} & V_{22} & \cdots & V_{k2}\end{bmatrix} \quad (4)$$

可以得到：

$$UV^{\mathrm{T}}=\begin{bmatrix}U_{11}V_{11}^{\mathrm{T}} & & & U_{11}V_{12}^{\mathrm{T}}\\ & U_{21}V_{21}^{\mathrm{T}} & & U_{21}V_{22}^{\mathrm{T}}\\ & & \ddots & \vdots\\ & & U_{k1}V_{k1}^{\mathrm{T}} & U_{k1}V_{k2}^{\mathrm{T}}\\ U_{12}V_{11}^{\mathrm{T}} & U_{22}V_{21}^{\mathrm{T}} & \cdots & \sum_{i=1}^{k}U_{i2}V_{i2}^{\mathrm{T}}\end{bmatrix}$$

$$=\begin{bmatrix}D_1 & & & C_1\\ & \ddots & & \vdots\\ & & D_k & C_k\\ R_1 & \cdots & R_k & kB\end{bmatrix} \quad (5)$$

观察分解结果发现，UV^{T} 矩阵和双边块对角矩阵的唯一区别就是双边的交叉块 B，UV^{T} 矩阵相比双边块对角矩阵 $\tilde{X}$，UV^{T} 矩阵的双边交叉块为 $\tilde{X}$ 矩阵的双边交叉块乘以对角块个数。根据这个思想，本文提出了基于双边块对角矩阵的矩阵分解算法，该算法的思想为将原始矩阵通过执行图分割算法排列为双边块对角矩阵后(见式(1))，再将双边块对角矩阵排列成块对角矩阵的形式(见式(2))，通过对子对角块矩阵进行矩阵分解，并对重复预测的子矩阵求平均，去恢复原始矩阵，具体算法将在 4.3 节详细描述。

4.2 得到规模相似的对角块矩阵

在实际的并行化矩阵分解中，通常一个线程对一个子矩阵进行分解，这就要求每个子矩阵的规模应该尽可能地相似，因为若存在规模相差较大的两个子矩阵，对规模较小的子矩阵进行分解的线程必定要等待规模较大的子矩阵进行分解的线程完成后才能对原始矩阵进行恢复，这会大大影响算法效率。为了得到规模尽可能相似的子对角块矩阵，本文提出算法 1，用于产生规模相似的 k 个对角块，这 k 个对角块可以对应到式(2)中，其算法描述如算法 1 所示。

算法 1 Balanced-BDF

Require：原始评分矩阵 $\mathbf{X}$，对角块矩阵目标密度 ρ，当前块对角矩阵中总对角块个数 k

Ensure：由双边块对角矩阵拼接而成的块对角矩阵 $\tilde{\mathbf{X}}$

```
1. function BDF(X,ρ,k)
2.   ρ'←ρ̄(X̃1,X̃2,X̃3,…,X̃k)
3.   if ρ'>ρ then
4.     return X̃
5.   else
6.     [D1,D2,D3,…,Dk]←SortByArea([D1,D2,D3,…,Dk])
7.     for i=1→k do
```

```
8.      [D_i^1, D_i^2] ← MetisNodeBDF(D_i)
9.      X̃_i 重新排列成 X̃_i^1, X̃_i^2
10.     if ρ̄(X̃_1, ..., X̃_i^1, X̃_i^2, ..., X̃_k) > ρ' then
11.       for j = k → i do
12.         X̃_{j++} = X̃_j
13.       end for
14.       X̃_i = X̃_i^1, X̃_{i+1} = X̃_i^2
15.       X' ← (X̃_1, X̃_2, ..., X̃_{k+1})
16.       Balanced-BDF(X', ρ, k+1)
17.     end if
18.   end for
19.   return X̃
20. end if
21. end function
```

基于点割集的分割算法是借助著名的开源图分割工具Metis,对双边块对角矩阵中的对角块进行分割,算法通过控制对角块的密度 ρ 来控制分割的进行,类似于式(2),如果这些对角块拼接起来的对角块子矩阵密度能得到提高,则将对角块拼接起来,并进行递归,直至所有的对角块平均密度达到预期要求后停止。

算法 1 在对对角块进行分割前,先对对角块的大小进行了一个降序的排列,在分割时也优先将面积较大的对角块先进行分割,基于面积的分割方法能保证分解出来的矩阵具有相似的面积,即子矩阵拥有相似的规模,方便后续的并行化处理。

4.3 基于对角块矩阵的矩阵分解

基于式(2)所展示的结构,根据公认的基于机器学习的矩阵分解框架[22],执行分解算法 $P=(f,D_W,C,R)$,利用每个子对角块 $\boldsymbol{X}_i$ 的分解结果 $\boldsymbol{U}_i,\boldsymbol{V}_i$ 去近似原始矩阵,分解算法每一部分的定义如下:

(1)基于矩阵分解算法中的预测函数 f,对于对角线上每个子对角块有 $\boldsymbol{X}_{ii}\approx f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}})=\boldsymbol{U}_i\boldsymbol{V}_i^{\mathrm{T}}$,即矩阵分解算法中的预测函数为 $f(x)=x$。对于式(2)所示的块对角矩阵,当 $i\neq j$ 时,$X_{ij}=0$,但是可以用对角线上的分解结果对非对角线上的零矩阵块进行预测填充,即 $\boldsymbol{X}_{ij}\approx f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}})=\boldsymbol{U}_i\boldsymbol{V}_j^{\mathrm{T}}$。

(2)D_W 为损失函数,$D_W(X,f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}))\geqslant 0$ 用于表示经过预测函数 $f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}})$ 预测后矩阵与原始矩阵的误差,其中 $\boldsymbol{W}$ 为数据权重矩阵,是损失函数的自变量。

对于本文算法,对于整个矩阵 $\boldsymbol{X}$,令损失函数等于每个预测点的损失值的平方和与其权重的乘积,损失值用真实值与预测值之差来表示,即:

$$\begin{aligned}D_W(\boldsymbol{X},f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}))&=\sum_{i,j}D_{W_{ij}}(\boldsymbol{X}_{ij},f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}})_{ij})\\&=\sum_{i,j}W_{ij}\times(\boldsymbol{X}_{ij}-f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}})_{ij})^2\end{aligned}\tag{6}$$

在实际的推荐系统中,要处理的矩阵通常都是十分稀疏的矩阵,包含了大量的未观测值,为了方便处理,常置这些未观测值为 0,但并不意味着用户对它们的评分为 0。在本文提出的算法中,对于未观测值,令其权重 $\boldsymbol{W}=0$,对于本文提出的对角块矩阵中的非对角线上的对角块,有 $D_{W_{ij}}(\boldsymbol{X}_{ij},f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}})_{ij})=0(i\neq j)$。

(3)C 为分解因子的约束。面向常见的含有显式评分的系统中,评分数一般都为正数,因此在本文算法中,硬性约束需要满足非负性约束,即分解结果 $\boldsymbol{U},\boldsymbol{V}$ 矩阵中不包含负数元素。

(4)R 为正则化项,为了防止过拟合问题,在本文提出的矩阵分解算法中,采用的是最常见的 ℓ_p-*norm* 正则化项:

$$\begin{aligned}R(\boldsymbol{U},\boldsymbol{V})&=\lambda_{\boldsymbol{U}}\ \|\boldsymbol{U}\|_p^p+\lambda_{\boldsymbol{V}}\|\boldsymbol{V}\|_p^p\\&=\sum_{i=1}^{k}(\lambda_{\boldsymbol{U}}\|\boldsymbol{U}_i\|_p^p+\lambda_{\boldsymbol{V}}\|\boldsymbol{V}_i\|_p^p)\\&=\sum_{i=1}^{k}R(\boldsymbol{U}_i,\boldsymbol{V}_i)\end{aligned}\tag{7}$$

具体算法采用的是 $p=2$ 的正则化项,即最著名的 Frobenius 正则化,采取该正则化项可以使分解结果更具有鲁棒性。

基于上述的几个参数,将矩阵分解问题转化为机器学习上的求解优化目标,对每一个子矩阵 $\boldsymbol{X}_i(1\leqslant i\leqslant k)$ 进行分解,便可以得到分解结果集合 $\boldsymbol{U}=\{\boldsymbol{U}_1,\boldsymbol{U}_2,\cdots,\boldsymbol{U}_k\},\boldsymbol{V}=\{\boldsymbol{V}_1,\boldsymbol{V}_2,\cdots,\boldsymbol{V}_k\}$,其优化目标如下:

$$\arg\min_{(\boldsymbol{U},\boldsymbol{V})\in C}[(\boldsymbol{X}-f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}))^2+\lambda_{\boldsymbol{U}}\ \|\boldsymbol{U}\|^2+\lambda_{\boldsymbol{V}}\ \|\boldsymbol{V}\|^2)]\tag{8}$$

在每轮迭代的过程中,$\boldsymbol{U}_i,\boldsymbol{V}_j$ 的每轮迭代更新可以表示为:

$$\boldsymbol{U}_i\leftarrow\boldsymbol{U}_i+W_{ij}\times\gamma\times(D_{W_{ij}})\times\boldsymbol{X}_{ij},f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}})_{ij}\times\boldsymbol{V}_j-\lambda\times\boldsymbol{U}_i\tag{9}$$

$$\boldsymbol{V}_j\leftarrow\boldsymbol{V}_j+W_{ij}\times\gamma\times(D_{W_{ij}})\times\boldsymbol{X}_{ij},f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}})_{ij}\times\boldsymbol{U}_i-\lambda\times\boldsymbol{V}_j\tag{10}$$

4.4 基于对角块矩阵的矩阵分解

本文提出的基于双边块对角矩阵的矩阵分解算法的流程总结如下:先把原始的稀疏矩阵通过迭代转换为多层双边块对角矩阵,再按照类似于式(2)的形式排列为块对角矩阵,对子对角块矩阵执行矩阵分解算法,并通过分解结果去对原始矩阵进行近似。对于原始的稀疏评分矩阵 $\boldsymbol{X}$,对它进行矩阵预测的具体实现流程为:

(1)对 $\boldsymbol{X}$ 矩阵执行基于社区发现的点割集图分割算法,将其转换为双边块对角矩阵 $\boldsymbol{X}'$。

(2)双边块对角矩阵调节为式(2)所示的块对角矩阵 $\widetilde{\boldsymbol{X}}$,$\widetilde{\boldsymbol{X}}_{I_*\sim J_*}$ 表示块对角矩阵的子矩阵。

(3)对各个子块对角矩阵 $\widetilde{\boldsymbol{X}}_{I_*\sim J_*}$ 执行矩阵分解算法 P,得到分解结果$(\boldsymbol{U}_i,\boldsymbol{V}_i)$,通过对分解结果使用预测函数 f,得到近似的分解结果 $\widetilde{\boldsymbol{X}}_i^*$,即:

$$\widetilde{\boldsymbol{X}}_i\approx f(\boldsymbol{U}_i\boldsymbol{V}_i^{\mathrm{T}})\overset{\triangle}{=}\widetilde{\boldsymbol{X}}_i^*\overset{\triangle}{=}\widetilde{\boldsymbol{X}}_{ii}^*\tag{11}$$

对角线上非零的子对角块 $\widetilde{\boldsymbol{X}}_{ii}$ 的近似已得出。

(4)根据矩阵分解算法所得到的分解结果 $\boldsymbol{U}=\{\boldsymbol{U}_1,\boldsymbol{U}_2,\cdots,\boldsymbol{U}_k\},\boldsymbol{V}=\{\boldsymbol{V}_1,\boldsymbol{V}_2,\cdots,\boldsymbol{V}_k\}$,对块对角矩阵中非对角线上的零块进行近似。

$$\widetilde{\boldsymbol{X}}_{ij}\approx f(\boldsymbol{U}_i\boldsymbol{V}_j^{\mathrm{T}})\overset{\triangle}{=}\widetilde{\boldsymbol{X}}_{ij}^*\tag{12}$$

至此,块对角矩阵所有值的预测填充工作都已完成,整个块对角矩阵的近似已得出。

(5)对重复预测的矩阵 $\widetilde{\boldsymbol{X}}_{ij}$ 求平均,得到近似的双边块对角矩阵。以式(1)中的 R_{12} 块为例子,R_{12} 由原始双边块对角矩阵 $\boldsymbol{X}'_{I_{B1}\sim J_{12}}$ 构成,在近似原始矩阵过程中,该子块在 $\widetilde{\boldsymbol{X}}_{12},\widetilde{\boldsymbol{X}}_{22}$ 子矩阵块中共被重复预测两次(见式(2)),因此在近似原始双边块对角矩阵的 $\boldsymbol{X}^*_{I_{B1}\sim J_{12}}$ 的值为:

$$\boldsymbol{X}^*_{I_{B1}\sim J_{12}}=\frac{1}{2}(\widetilde{\boldsymbol{X}}^{*(12)}_{I_{B1}\sim J_{12}}+\widetilde{\boldsymbol{X}}^{*(22)}_{I_{B1}\sim J_{12}})\tag{13}$$

同理，对块对角矩阵中每个 $\widetilde{\boldsymbol{X}}_{I\sim J}$ 做同样的工作，便可得到近似的原始矩阵，即：

$$\boldsymbol{X}^{*}=\{\boldsymbol{X}^{*}_{I^{*}\sim J^{*}}\} \tag{14}$$

至此，矩阵分解及近似恢复原始矩阵的工作完成。

4.5 性能分析

Metis 中基于点割集的图分割算法的复杂度为 $O(n)$[23]，n 为图中边的数量，对应到评分矩阵中即非零值的个数。对于本文算法，假设最终产生 k 个对角块，通常 $k\ll n$，递归算法的递归树深度为 $O(\lg k)$，因此本文算法的复杂度为 $O(n\lg k)$。

5 实验结果

本节将本文算法与最新先进的推荐算法进行对比，将评估标准作为各种推荐算法的性能参考。实验的实验机的处理器为 AMD R3 3300X 3.8GHz CPU，实验机的内存为16GB，实验机的操作系统为 Windows 10，实现基线算法以及本文算法编程语言为 C++，编程所使用的集成开发软件为 Visual Studio 2015。

本节将先介绍实验所使用的数据集信息，然后介绍对比算法与评估标准，得出对比实验的结果后，进一步探讨影响本文提出的推荐算法精度的参数。

5.1 使用的数据集信息

对于对比实验，本文基于以下的公开的真实数据集进行研究：Movielens-1m，Book-crossing。其中，Movielens-1m 是 GroupLens 研究组对电影评分进行收集而来的数据集[1)]，是推荐系统研究中最广泛用于评估推荐系统的性能的数据集，Movielens-1m 数据集包含了 6 040 个不同的用户对约 3 900 部电影的大约 100 万个评分，评分的区间为 1～5 分。Book-crossing 是根据 bookcrossing.com 的数据编写的图书评分数据集[2)]。该数据集包含了大约 27 万个独立用户对约 27 万本书的 110 万次评分，评分区间为 1～10 分。同时，Book-crossing 是目前个性化推荐系统研究中最不密集的数据之一。两个数据集信息如表 1 所列。

表 1 实验中的数据集信息

Table 1 Statistics of two datasets

Datasets	Users	Items	Ratings	Density/%	Rating range
Movielens-1m	6040	3952	1 000 209	4.2600	1～5
Book-crossing	278 858	271 379	1 149 780	0.0014	1～10

上述两个数据集的规模、密度均有较大的差别，能很好地探究本文算法在不同场景下的表现。

5.2 算法及评价指标的选择

本文选择了以下 4 个矩阵分解算法作为基线算法，用于对比研究基于双边块对角矩阵的矩阵分解的精度。

(1)SVD[24]：奇异值分解，采用的是 Simon Funk 在 Netflix 大赛期间提出的算法，该算法采用的是 Frobenius 范数。

(2)SVD++[25]：考虑用户的历史行为对用户评分预测的影响的 SVD 算法的扩展。

(3)PMF[26]：概率矩阵分解算法，建立用户行为概率模型，并根据这个模型进行预测。

(4)NMF[27]：非负矩阵分解算法，同样采用的是 Frobenius 范数。

在对矩阵评分精度预测的研究上，本文选用的评价指标为均方根误差(RMSE)，假设测试集中共有 N 个用户 u 对物品 i 的评分 r_{ui}，若经过协同过滤算法，物品的预测值为 $\hat{r}_{ui}$，则 RMSE 可以表示为：

$$RMSE=\sqrt{\frac{\sum_{i=1}^{N}(r_{ui}-\hat{r}_{ui})^{2}}{N}} \tag{15}$$

本文采用五折交叉验证进行实验，将数据集分为 5 个部份，每次实验选择其中 4 份作为训练集，1 份作为测试集，共重复 5 次实验，每次选取的训练集都不同，通过对 5 次实验计算出来的 RMSE 值求平均，将其作为最终的 RMSE 值。

5.3 算法的分析

5.3.1 不同目标密度对对角块数量的影响

根据本文算法，生成对角块的数量是由用户设置的密度 ρ 所决定的。图 3 和图 4 给出了在两个数据集上设置的密度 ρ 和最终形成的对角块个数的关系。

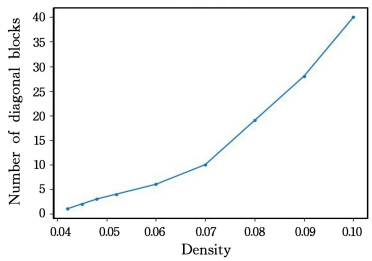


图 3 Movielens-1m 在不同密度下对角块个数

Fig. 3 Number of diagonal blocks in Movielens-1m under different densities

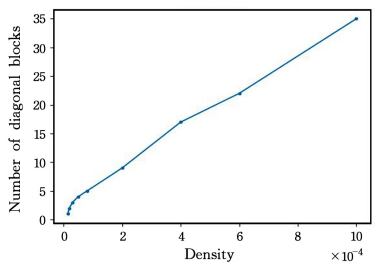


图 4 Book-crossing 在不同密度下对角块个数

Fig. 4 Number of diagonal blocks in Book-crossing under different densities

根据图 3、图 4 可以得知，随着设置的密度的增大，对角块的个数也会增多，假设设置的目标密度小于原始矩阵的密度，那么就可以将原始矩阵看作是一个对角块，生成对角块个数为 1。但是，若目标密度设置得过大，将会生成很多小规模的矩阵，这是因为随着过多对角块的生成，这些对角块的规模将会变得非常小，以 Movielens-1m 数据集生成的对角块为例，当设置的目标密度为 0.1 时，生成了 40 个平均用户个数仅为 201 个，而平均物品个数也只有 2 881 个的对角块矩阵，两个数据集在不同密度下产生的对角块中的平均用户数量以及物品数量如图 5 和图 6 所示。

1) http://grouplens.org/datasets/movielens

2) http://www2.informatik.uni-freiburg.de/~cziegler/BX/

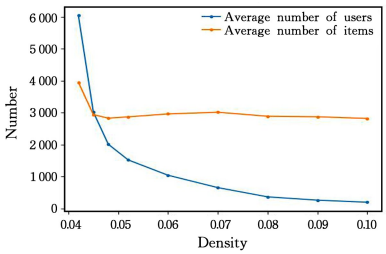


图 5 Movielens-1m 在不同密度下每个对角块平均用户数和平均物品数

Fig. 5 Average number of users and items per diagonal block in Movielens-1m under different densities

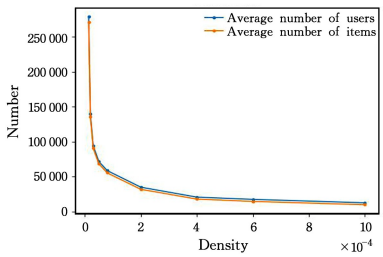


图 6 不同密度下每个对角块平均用户数和平均物品数

Fig. 6 Average number of users and items per diagonal block in Book-crossing under different densities

过多的零散小规模矩阵会十分影响矩阵分解算法对用户偏好的正确建模，后面将会设计实验验证，对角块并非越多越好。在实际的应用中，同样不需要这么多的对角块，为了并行化的处理，通常生成对角块的个数为计算机 CPU 的内核数（或整数倍），通过一条线程负责处理一个对角块即可提升矩阵分解的效率。

5.3.2 本文算法与基线算法的对比

对于矩阵分解算法 $P=(f,D_W,C,R)$ 来说，矩阵的分解因子数量 r 对评分预测的精度影响十分重大，若分解因子的数量太小，原始矩阵构建的用户偏好和物品属性模型过于简单，不能很好地去拟合原始矩阵，若分解因子数量过大，生成模型耗费的时间将会很长，并且也会带来过拟合的后果，最终的预测精度也会下降，因此选取一个恰当的分解因子数量十分重要。

将 Movielens-1m 的数据集作为实验数据，对数据集的评分矩阵执行目标密度 $\hat{\rho}=0.052$ 的双边块对角化算法，原始矩阵将产生 4 个对角块，4 个对角块的信息如表 2 所列。

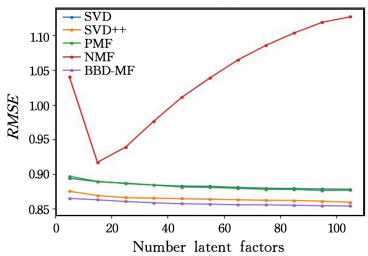
表 2 4 个对角块的信息

Table 2 Information of four diagonal blocks

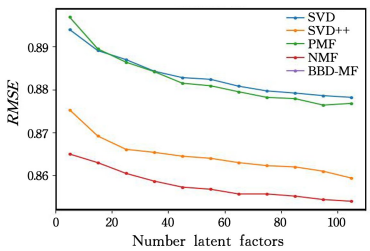
	X_1	X_2	X_3	X_4
Users	1 559	1 264	1 413	1 851
Items	2 440	2 766	3 091	3 590
Ratings	181 975	163 176	227 915	361 832
Density	0.047 8	0.046 7	0.052 2	0.054 5

为考察不同分解因子数量下，4 种基线算法以及本文提出的基于双边块对角矩阵的矩阵分解算法预测精度的区别，本文直接在原始矩阵上执行 4 种基线算法，计算并记录其 RMSE 值。作为对比，对产生的 4 个对角块并行执行本文提出的矩阵分解算法，计算并记录根据每个子对角块的分解结果去近似的矩阵的 RMSE 值。

本文将分解因子数量的范围设置为 5～105，每次实验的步长为 10。其中，SVD，SVD＋＋，NMF 算法的正则化项系数设置为 $\lambda=0.065$，PMF 算法的正则化项系数设置为 $\lambda_U=\lambda_V=0.002$，对于本文提出的矩阵分解算法，与 SVD 等算法具有相似之处，因此正则化系数也选择为 $\lambda=0.065$，五折交叉验证的实验结果如图 7 和图 8 所示。



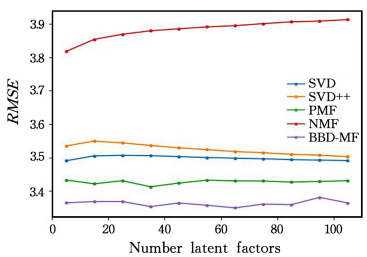
(a)5 种算法的对比



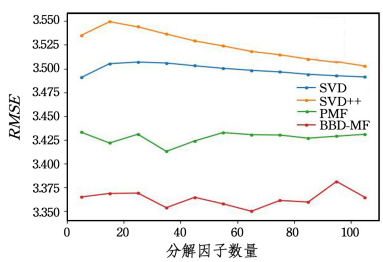
(b) 性能相近算法对比

图 7 Movielens-1m 在不同分解因子数量下 5 种算法的 RMSE

Fig. 7 RMSE of Movielens-1m for five algorithms with different numbers of lantent factors



(a)5 种算法的对比



(b) 性能相近算法对比

图 8 Book-crossing 在不同分解因子数量下 5 种算法的 RMSE

Fig. 8 RMSE of Book-crossing for five algorithms with different numbers of lantent factors

对于 Book-crossing 数据集，类似于 Movielens-1m 数据集，本文采取同样的算法，控制目标密度为 0.000 15，可以得到 8 个对角块，对这 8 个对角块采取本文算法，记录下预测的 RMSE 值，并与直接在原始矩阵上执行基线算法的结果进行比较，比较结果如图 8 所示。

在两个数据集上的实验结果如图 7、图 8 所示，5 条曲线分别为 4 种基线算法（SVD，NMF，SVD＋＋，PMF）以及本文提出的 BBD-MF 算法，实验结果表明，本文提出的算法在预测精度上均比 4 种基线算法有提升，其中在大规模稀疏数据集 Book-crossing 中表现更为突出。能得到这样的实验结果

是基于以下原因:1)算法中抽取子矩阵较原始的评分矩阵密度更高,在相对密集的子矩阵上执行协同过滤算法进行预测,能得到更高的预测结果;2)子矩阵是根据社区发现的图分割算法进行划分的,因此同一个子矩阵中的用户具有更相似的物品偏好,这对预测结果也会有提升。

同时,通过观察在 Movielens-1m 上执行 5 种算法得到的实验结果也能够发现,本文提出的基于双边块对角矩阵的矩阵分解算法在分解因子数量较小时就能得到相对较好的预测精度。这是因为对于 4 种基线算法来说,当分解因子数量较小时,模型过于简单,不足以去拟合庞大的原始矩阵,但对于本文提出的基于双边块对角矩阵的矩阵分解算法而言,算法首先得到数个对角块,矩阵分解后的预测工作是基于这些对角块的,对角块的规模相比原始矩阵要小得多,因此,即使分解因子数量较小,简单的模型也能很好地去拟合对角块矩阵。选择较小的分解因子就能获得相对较好的预测效果,这意味着能节省大量的建模时间,提高推荐算法的效率。

5.3.3 目标密度设置对预测精度的影响

本节将讨论目标密度设置对预测精度的影响,通过前面的实验已知在合理设置目标密度的前提下,对角块个数总是与目标密度成正比的,但是在实际的系统中,往往不需要过多的子矩阵。本节实验对 Movielens-1m 以及 Book-crossing 数据集中的评分矩阵设置不同的目标密度,对其执行双边块对角化算法,并产生对角块。根据 5.3.2 节的实验可知,在 Movielens-1m 数据集上,当分解因子数量为 65 时,预测的精度就趋于平稳,预测的精度也可以被接受;在 Book-crossing 数据集上分解因子个数为 35 时,预测的精度相对较好。因此,对于 Movielens-1m 数据集,固定分解因子数量为 65,对 Book-crossing 数据集固定分解因子数量为 35,并对这两个数据集产生的对角块执行文中提出的矩阵分解算法,正则化系数的设置与 5.3.2 节相同,其结果如图 9 和图 10 所示。

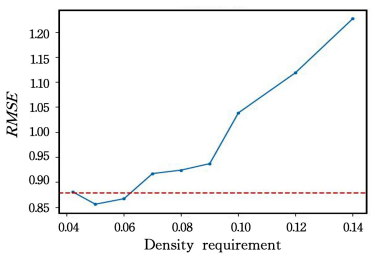


图 9 Movielens-1m 数据集在不同对角块密度下 BBD-MF 算法的 RMSE

Fig. 9 RMSE of BBD-MF algorithm for Movielens-1m on different densities of the diagonal block

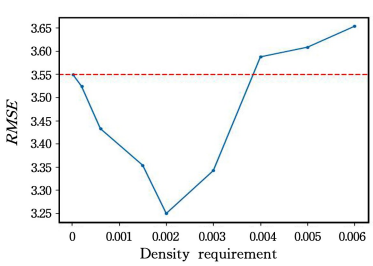


图 10 Book-crossing 数据集在不同对角块密度下 BBD-MF 算法的 RMSE

Fig. 10 RMSE of BBD-MF algorithm for Book-crossing on different densities of the diagonal block

图 9、图 10 中初始点为评分矩阵的原始密度下的 RMSE,即不对矩阵进行分割,执行本文提出的矩阵分解算法,并进行预测的 RMSE。为了直观比较,根据初始点的值绘制虚线,虚线表示不对矩阵进行分割直接进行矩阵分解的均方根误差。从图 9、图 10 中可以得知,RMSE 一开始随着目标密度的提升而降低,但是当目标密度设置得过大时,随着对角块的增多,RMSE 反而会比初始矩阵更大,矩阵的规模过小,将会大大影响推荐算法对用户偏好以及物品属性的建模。上述实验结果表明,对角块并不是越多越好,在实际的应用中,应选择合适的密度,产生合适数量的对角块进行预测。

5.3.4 并行化效率分析

在此部分,将对本文算法的并行化效率进行分析,实验的实验机核心数为 4 核 8 线程的 CPU,因此本次实验通过控制输入的密度,将原始矩阵分解为 8 个对角块,在 4 个物理内核上运行 8 个线程,每个线程对一个对角块进行矩阵分解算法,并记录形成对角块矩阵的时间与通过分解结果来预测原始矩阵所用的时间之和。对于基线算法,直接对原始矩阵执行 4 种基线算法,记录每种算法分解矩阵并完成预测所需的时间。算法的分解因子个数以及正则化项参数均与 5.3.3 节的实验保持一致,算法执行完成后,结果如表 3 所列。

表 3 在两个数据集下算法耗时的对比

Table 3 Computational time of all algorithm on different dataset

Algorithm	Movielens-1m	Book-crossing
SVD	17.6 s	23.4 s
NMF	43.3 s	76.5 s
SVD++	27.4 min	1.05 h
PMF	18.42 s	22.5 s
BBD-MF	12.5 s	20.7 s

本文算法通过将原始矩阵分解为若干个规模相似的对角块,每个对角块都能独立执行矩阵分解算法,这使得矩阵分解算法的并行化效率大幅度提升。同时,当整体的数据发生变化后,不需要对整个矩阵进行重复的分解,只需要对发生改变的子矩阵进行重新分解即可,通过重新分解得来的分解结果去预测原始矩阵。并且对于越稀疏的矩阵,本文算法对比基线算法的提升越大,通过将原始矩阵分解成若干个对角块,使得不需要用很复杂的分解方法去训练模型,模型也会有较好的效果,这对实际系统中处理大规模稀疏矩阵有着重大的意义。

结束语 为了缓解推荐系统中数据稀疏性带来的种种问题,本文受社区发现算法启发,考虑到同一社区中的用户具有相似的兴趣偏好,本文从矩阵内部结构出发,使用基于社区发现的点割集图分割算法,对原始矩阵进行划分,可得到双边块对角矩阵,通过将双边块对角矩阵的双边和对角块拼接,能得到密度比原始矩阵更高并且可以并行进行分解的子矩阵。

本文实验部分,基于 Movielens-1m 数据集以及大规模稀疏数据集 Book-crossing 进行了对比实验。实验结果表明,通过设置合理的密度参数,就能生成合适数量的子对角块矩阵,通过合理设置正则化项、分解因子数量等参数,对比基线算法,本文算法能缓解数据稀疏性带来的问题,可以得到更好的预测精度,同时算法耗时也会更短,特别是在大规模、极其稀疏的数据集下表现更优。

本文在缓解推荐系统冷启动问题上还有非常大的提升空间,这有待更进一步开展实验进行详细的研究。同时,在未来

的研究中，拟将本文算法融入分布式大数据框架之中，以验证本文算法在处理特大规模、极其稀疏数据时的优越性。

参 考 文 献

[1] BOBADILLA J，ALONSO S，HERNANDO A. Deep learning architecture for collaborative filtering recommender systems[J]. Applied Sciences，2020，10(7)：2441.

[2] WU L，GE Y，LIU Q，et al. Modeling the Evolution of Users' Preferences and Social Links in Social Networking Services[J]. IEEE Transactions on Knowledge and Data Engineering，2017，29(6)：1240-1253.

[3] BOBADILLA J，ORTEGA F，GUTIERREZ A，et al. Classification-based Deep Neural Network Architecture for Collaborative Filtering Recommender Systems[J]. International Journal of Interactive Multimedia & Artificial Intelligence，2020，6(1)：68-77.

[4] FENG C，LIANG J，SONG P，et al. A fusion collaborative filtering method for sparse data in recommender systems[J]. Information Sciences，2020，521：365-379.

[5] LI D S，CHEN C，LV Q，et al. An algorithm for efficient privacy-preserving item-based collaborative filtering[J]. Future Generation Computer Systems，2016，55：311-320.

[6] SAMMEL M D，RYAN L M. Latent variable models with fixed effects[J]. Biometrics，1996，52：650-663.

[7] DEERWESTER S，DUMAIS S T，FURNAS G W，et al. Indexing by latent semantic analysis[J]. Journal of the American society for information science，1990，41(6)：391-407.

[8] LEE D D，SEUNG H S. Learning the parts of objects by non-negative matrix factorization[J]. Nature，1999，401(6755)：788-791.

[9] MNIH A，SALAKHUTDINOV R R. Probabilistic matrix factorization[J]. Advances in Neural Information Processing Systems，2007，20：1257-1264.

[10] NAJAFABADI M K，MAHRIN M N，CHUPRAT S，et al. Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on implicit data [J]. Computers in Human Behavior，2017，67(FEB.)：113-128.

[11] WANG Y G，SONG Z Z，XIAO C L. Collaborative filtering recommendation algorithm based on improved clustering and matrix factorization[J]. Journal of Computer Applications，2018，38(4)：1001-1006.

[12] ABBASI-MOUD Z，VAHDAT-NEJAD H，SADRI J. Tourism recommendation system based on semantic clustering and sentiment analysis[J]. Expert Systems with Applications，2021，167：114324.

[13] AALDERING L J，LEKER J，SONG C H. Recommending untapped M&A opportunities：A combined approach using principal component analysis and collaborative filtering[J]. Expert Systems with Applications，2019，125：221-232.

[14] NGUYEN T T. On the Edge and Cloud：Recommendation Systems with Distributed Machine Learning[C]// 2021 International Conference on Information Technology(ICIT). IEEE，2021：929-934.

[15] YU H F，HSIEH C J，SI S，et al. Scalable coordinate descent approaches to parallel matrix factorization for recommender systems[C]// 2012 IEEE 12th International Conference on Data Mining. IEEE Computer Society，2012：765-774.

[16] LIU Q Q，LUO Y L，WANG Y F，et al. Hybrid Recommendation Algorithm Based on SVD Filling[J]. Computer Science，2019，46(S1)：468-472.

[17] SU Y，ZHOU K，ZHANG X，et al. A parallel multi-objective evolutionary algorithm for community detection in large-scale complex networks[J]. Information Sciences，2021，576：374-392.

[18] OSABA E，DEL SER J，CAMACHO D，et al. Community detection in networks using bio-inspired optimization：Latest developments，new results and perspectives with a selection of recent meta-heuristics[J/OL]. Applied Soft Computing，2020，87. https://doi.org/10.1016/j.asoc.2019.106010.

[19] FORTUNATO S，HRIC D. Community detection in networks：A user guide[J]. Physics reports，2016，659：1-44.

[20] ZHAO W J，ZHANG F B，LIU J L. Review on community Detection in Complex Networks[J]. Computer Science，2020，47(2)：10-20.

[21] GLIGORIJEVIĆ V，PANAGAKIS Y，ZAFEIRIOU S. Non-negative matrix factorizations for multiplex network analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2018，41(4)：928-940.

[22] SINGH A P，GORDON G J. Relational learning via collective matrix factorization[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery. 2008：650-658.

[23] KARYPIS G，KUMAR V. A fast and high quality multilevel scheme for partitioning irregular graphs[J]. SIAM Journal on Scientific Computing，1998，20(1)：359-392.

[24] KOREN Y，BELL R，VOLINSKY C. Matrix factorization techniques for recommender systems[J]. Computer，2009，42(8)：30-37.

[25] KOREN Y. Factorization meets the neighborhood：a multifaceted collaborative filtering model[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery. 2008：426-434.

[26] SALAKHUTDINOV R，MNIH A. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo[C]// Proceedings of the 25th International Conference on Machine Learning. Association for Computing Machinery. 2008：880-887.

[27] LUO X，ZHOU M C，XIA Y B，et al. An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems[J]. IEEE Transactions on Industrial Informatics，2014，10(2)：1273-1284.



HE Yi-chen，born in 1998，postgraduate，is a student member of China Computer Federation. His main research interests include recommendation system and data mining.



GU Wan-rong，born in 1982，Ph.D，assistant professor. His main research interests include recommendation system，information retrieval and big data processing.

DOI: 10.20266/j.math.22-0133

管理科学

居民消费价格指数、工业增加率与股票指数波动关系分析——基于 GARCH-MIDAS 模型

谢贤芬[1], 何亦琛[2], 毛宜军[3], 古万荣[3,*], 张子烨[2], 黄天睿[4]

(1. 暨南大学 经济学院, 广东 广州 510632)
(2. 华南理工大学 计算机科学与工程学院, 广东 广州 510641)
(3. 华南农业大学 数学与信息学院, 广东 广州 510642)
(4. 广东第二师范学院 计算机学院, 广东 广州 510303)

摘 要: 为探究宏观经济因素与股市波动之间的关系, 选取上证 180 指数与中证 100 指数为研究对象, 将居民消费价格指数和工业增加率作为研究变量, 基于多因子 GARCH-MIDAS 模型, 研究宏观经济变量对股票价格长短期波动的影响. 结果表明: 模型能够很好地拟合股市的波动性, 能较好地反映宏观经济变量与股市波动性之间的相关性; 两个变量均对股市长期波动产生正向影响, 均能显著影响未来股票市场波动. 两者对股市波动短期影响则各有不同, 工业增加率的影响不明显, 而股市短期波动对居民消费价格指数的波动非常敏感.

关键词: 股市波动率; GARCH-MIDAS; 混频数据; 长短期波动

1 引 言

股票市场的剧烈波动会对整个经济运行产生巨大的影响, 甚至导致金融危机或经济危机, 我国经济正从高速发展向高质量发展转型, 股票市场的波动与我国实体经济的发展也息息相关, 因此完善金融体系以及股票市场, 保证股票市场的健康发展是经济持续发展中重要的一环. 同时, 资产的波动性研究也是金融资产定价的焦点问题, 监管机构及国内学者都高度重视对股票价格波动的研究. 对国内来说, 我国股票市场发展较晚, 市场的有效性不充足, 股市的内在机制与规章制度仍不完善, 股市容易发生较大幅度的波动, 因此研究宏观经济与股市的波动关系, 利于政府完善市场监管手段, 为我国通过宏观调控防止股市剧烈波动提供建议与参考, 具有非常重大的理论与实际意义.

本文利用 GARCH-MIDAS 混频数据模型, 将选取的股市数据上证 180 指数和中证 100 指数的波动分解为长期成分以及短期成分, 从多因素出发研究居民消费价格指数与工业增加

收稿日期: 2022-02-22
资助项目: 广东省自然科学基金面上项目 (2022A1515011489); 全国统计科学研究项目 (2023LY016); 国家自然科学基金项目 (62472182)
* 通信作者

中国知网 https://www.cnki.net

率二个宏观经济变量对股市波动的影响, 并进一步验证模型的稳健性及适用性. 本文创新点在于:

1) 采用混频模型对股票指数波动进行研究, 能在不降低股票价格数据采样频率的前提下, 充分利用股票价格高频日数据, 将高频的股票价格数据与低频宏观经济数据相结合, 从而更精细地研究股票价格地波动情况.

2) 现有文献中涉及的宏观经济变量指标有经济增长、通货膨胀、经济景气指数、政策不确定性、人均收入等, 暂未见有将工业增加率作为外生变量进行分析, 而工业增加率是国民经济核算的一项基础指标, 能够反映整个国家的资金流量. 本文首次将工业增加率指标纳入到 GARCH-MIDAS 模型的构建中, 得到的结果将更具有说服力, 丰富了宏观经济因素与股票价格的实证研究.

2 文献回顾

导致股票市场波动的原因有短期的股市实现机制因素, 也有长期的宏观经济政策因素, 从长期来看, 宏观经济变量对股票市场波动的影响研究更为重要. 顾海峰和周亚勇 [1] 通过建立 GARCH 模型, 对中证 500 股指期货的推出对中国股票市场波动的长期影响与短期影响进行了研究, 研究结果表明, 中证 500 股指期货的推出对中国股票市场长短期影响各不相同, 从长期来看, 期货的推出对股市波动具有长期抑制效应, 从短期来看, 期货的推出对股市波动具有显著的助推效应. 刘庞庞 [2] 建立 ARMA-GARCH 与 TGARCH 模型, 对上证 50ETF 期权收益率的价格波动对现货市场波动的影响进行了研究, 实证结果表明期权上市后对现货市场价格波动的影响非常大, 存在严重的非对称性现象, 但是长期来看, 期权的上市会对现货市场起到稳定的作用.

由于宏观经济变量数据往往是月度、季度或年度数据, 而股票价格数据一般是日数据或更为高频的 5 分钟数据, 对于不同频率的数据, 传统的时间序列分析方法已不再适用, 而利用插值进行混频建模的方法的预测精度不佳, 也会带来信息虚增问题, 未得到太多的重视. 直接建立混频数据模型能有效避免插值带来的信息虚增等问题, 同时也很大程度上提高了预测精度, 在很多领域得到了广泛应用. 混频数据模型不对混频数据做任何处理, 而是充分利用原始数据的信息去建模. 这些模型提出的初衷大部分是应用于金融市场的波动预测, 随着模型方法研究的不断深入, 2005 年开始也应用于宏观经济领域中. 至今国内外也产生了大量的研究成果, Asgharian 等 [3] 使用 MIDAS 模型去分析宏观经济不确定性对股市波动长期成分的影响. 彭伟 [4] 运用了 AR, HAR 和 MIDAS 模型预测股票日收益风险价值并进行比较研究, 发现: 在预测银行日收益风险方面, HAR 和 MIDAS 预测效果要优于 AR 模型.

在新近研究中, 更多的是利用 GARCH-MIDAS 模型. 周德才等 [5] 基于月频金融数据, 构建了我国金融状况指数 (FCI) 以表征货币政策的不确定性. 使用 GARCH-MIDAS 模型实证分析我国股市波动的长短期成分及货币政策不确定性对其的影响. 结果表明: 货币政策不确定性的水平值对我国股市波动没有显著影响, 但其波动率则有显著的正向影响. 石强等 [6] 运用多因子 GARCH-MIDAS 模型研究了我国宏观经济与股市波动之间的关系. 研究结果表明: 社会消费品零售总额会对股市长期波动产生正向影响, 并且这种影响有逐渐增强的趋势. 利率与货币供给量则在不同经济发展阶段对股市波动的影响表现出一定的差异性, 这主要是与

中国知网 https://www.cnki.net

当时的宏观经济环境以及经济变量自身的特征有关. 夏婷和闻岳春 [7] 利用 GARCH-MIDAS 模型分析了经济不确定性 (宏观经济不确定性和经济政策不确定性) 对中国股市波动率的影响, 结果发现, 经济不确定性会影响中国股市的波动, 但强度有限, 且 A 股、B 股间会表现出差异. 尹力博和李勍[8] 通过构建人民币关注指数作为投资者关注指数, 并通过 GARCH-MIDAS 模型将人民币汇率价差的波动分解为长期成分与短期成分, 对投资者关注对人民币汇率价差波动的影响进行了研究, 研究结果表明投资者关注能够显著增大人民币汇率价差波动, 相比于波动的长期成分, 投资者关注对波动的短期成分影响更大. 王江涛等 [9] 通过引入一类自适应权重函数, 提炼出高频数据的信息, 并与低频数据相融合, 改进 GARCH 模型, 构架出了能更稳健、更准确预测波动率的混频 GARCH 模型.

李丹 [10] 选取中国三大分级基金指数系列的日收益率序列作为研究样本, 利用 GARCH 族模型研究分级基金波动收益的特征. 论文首先通过 AIC 准则确定不同类型分级基金的收益率的残差分布, 并确定最佳的 ARIMA-GARCH 模型, 在这个模型的基础上, 通过利用 GARCH-M 模型研究分级基金波动与收益的关系, 利用 TGARCH 模型和 EGARCH 模型研究分级基金波动的非对称性. 实证结果表明股票型分级基金指数对前期市场变化反应程度比债券型分级基金更为平缓, 股票型分级基金指数的条件方差的冲击的持续性要比债券型分级基金更大, 波动存在着长期记忆. Asgharian 等 [11] 利用 DCC-MIDAS 方法研究股票市场价格长期成分与债券相关性. 研究表明, 长期成分的相关性既受宏观金融变量 (历史以及预测) 的影响, 又受股市价格自身的影响. Ashwani 和 Sheera[12] 运用 GARCH-MIDAS 方法, 研究宏观经济变量在解释印度股市波动中的作用. 研究发现, 汇率、货币供应量、国库券利率等宏观经济变量对股市波动具有可预测的能力.

Yu 和 Huang[13] 研究经济政策不确定性对股票波动预测的影响, 运用 GARCH-MIDAS 模型, 将低频经济政策不确定性指数和高频股票收益率直接结合起来, 对股指波动进行建模和预测. 结果表明, 在 GARCH-MIDAS 模型中, 将实际波动率与经济政策不确定性相结合, 可以显著提高股指波动率的预测的准确性. Wang 等 [14] 将非对称性和极端波动效应相结合, 建立了预测股票波动的 GARCH-MIDAS 模型. 并发现在短期波动模型中引入了非对称和极端波动效应, 使得预测能力的提高更加明显. 张一锋等 [15] 提出了一种新的羊群效应测度指数, 运用 GARCH-MIDAS 模型验证了羊群效应测度指数对上证综指波动率的影响. 通过比较纳入新羊群效应指数与经济政策不确定性指数 (EPU) 的 GARCH-MIDAS 模型, 证实纳入新羊群效应指数的 GARCH-MIDAS 模型能更好地描述上证综指的长期成分, 并且能显著提高上证综指波动率的样本外预测精度. 同样, 利用 GARCH-MIDAS 模型研究股市波动率的文献还有[16,17].

3 混频 GARCH-MIDAS 模型

国内外对金融资产波动的研究总结出金融资产收益率波动具有波动聚集性、非对称性等共同特性, 大量的实证研究表明 GARCH 族模型能很好的拟合其波动性 [18], 成为目前金融计量中最为广泛应用的模型之一. 但 GARCH 模型不能处理混频数据, 为了更好利用 GARCH 类模型, Engle 和 Ghysels[19] 将 GARCH 模型与 MIDAS 混频抽样模型结合起来, 提出了能将波动分解为长期成分以及短期成分的 GARCH-MIDAS 模型, 模型能在不引入噪声的前提之

中国知网 https://www.cnki.net

下, 研究不同频率的变量之间的关系. 模型的短期成分由股市的历史波动所决定, 长期成分由宏观经济因素解释变量决定. 在该模型中, 收益率的方程可表示为

$$r_{i,t} = \mu + \sqrt{\tau_t g_{i,t}}\varepsilon_{i,t}, \quad \forall i = 1, 2, \cdots, N_t, \tag{1}$$

式中, $r_{i,t}$ 代表股市在 t 月的第 i 天的对数收益率, μ 通常为常数, 一般为 0, τ_t 是波动的长期成分, $g_{i,t}$ 是波动的短期成分. $\varepsilon_{i,t}$ 是服从 $\Phi_{i-1,t}$ 条件下的标准正态分布的随机扰动项, 即 $\varepsilon_{i,t}|\Phi_{i-1,t} \sim N(0,1)$, $\Phi_{i-1,t}$ 是在第 t 期时获得的前 $i-1$ 天的历史信息集, 对于短期成分 $g_{i,t}$, 假定其服从日度的 GARCH(1,1) 动态过程, 即

$$g_{i,t} = (1 - \alpha - \beta) + \alpha\frac{(r_{i-1,t} - \mu)^2}{\tau_t} + \beta g_{i-1,t}, \tag{2}$$

式中, α, β 参数衡量的是 GARCH 过程的效应, $\alpha + \beta < 1$ 但是接近 1, 对于长期成分 τ_t, 可以根据 MIDAS 回归得到的已实现波动率 RV_t 对其进行建模, 其方程如下所示:

$$\tau_t = m + \theta\sum_{k=1}^{K}\varphi_k(\omega_1, \omega_2)RV_{t-k}. \tag{3}$$

考虑到本文包含居民消费价格指数、以及工业增加率两个宏观变量, 为了消除这些数据的异方差带来的影响, 使得 GARCH-MIDAS 更好与宏观经济变量匹配, 保证长期成分始终为正数, 对于长期成分 τ_t 作对数处理, 即长期成分为

$$\log\tau_t = m + \theta\sum_{k=1}^{K}\varphi_k(\omega_1, \omega_2)RV_{t-k}, \tag{4}$$

式中, 已实现的波动率 RV_t 表示资产过去一段时间内的波动收益率, 参考郑挺国和尚玉皇 [20] 的方法, 以某一期中所有收益率的平方和来表示, 即

$$RV_t = \sum_{i=1}^{N_t} r_{i,t}^2. \tag{5}$$

$\varphi_k(\omega_1, \omega_2)$ 表示基于 Beta 函数构造的权重方程, 其方程如下:

$$\varphi_k(\omega_1, \omega_2) = \frac{(k/K)^{\omega_1 - 1}(1 - k/K)^{\omega_2 - 1}}{\sum_{j=1}^{K}(j/K)^{\omega_1 - 1}(1 - j/K)^{\omega 2 - 1}}. \tag{6}$$

为了保证滞后的宏观经济变量的权重是衰减的形式, 同时也为了简化参数的模型估计, 避免参数过多, 本文设置 $\omega_1 = 1$, 即单权重的 Beta 函数. 在式中, K 是低频变量中的最大滞后阶数, 由 AIC 与 BIC 信息准则来确定.

根据公式 (1) 至 (6) 的方程, 即可构建一个对数化的 GARCH-MIDAS 模型. 对于式 (3), 可以视其为具有潜在回归因子的回归模型, 因此, 这里采用极大似然估计法估计模型参数. 在本文中, 需要估计的 GARCH-MIDAS 模型参数包括 $\mu, \alpha, \beta, m, \theta, \omega$. 对数似然函数可表示为

$$LLF = -\frac{1}{2}\sum_{t=1}^{T}\left[\log(2\pi) + \log g_t(\Phi)\tau_t(\Phi) + \frac{(r_t - \mu)^2}{g_t(\Phi)\tau_t(\Phi)}\right], \tag{7}$$

在式子中, T 为总的周期数.

中国知网 https://www.cnki.net

4 数据预处理及基本统计分析

4.1 宏观经济变量的选取

通货膨胀是普遍存在的经济现象, 是衡量购买力与产出的指标, 若购买力大于产出, 则会引起物价的上涨, 从而使货币贬值, 投资者为了达到保值的目的, 会选择将资金投入股票市场, 股票的价格将会有所提高. 居民消费价格指数 (CPI) 是衡量通货膨胀的一个重要指标, 能反映与居民生活相关的商品价格的变动状况, 本文选择 2010 年 1 月至 2021 年 2 月的月度 CPI 数据作为宏观经济变量之一.

我国是制造业大国, 工业增加值是工业企业以货币形式表现的工业生产活动的最终成果, 反映了生产单位和部门对国内生产总值的贡献, 衡量着我国经济发展水平, 是国民经济核算的基础性指标, 是统计资金流量的基础. 本文选取的工业增加值率 (IP) 代表工业生产新增价值与整体生产价值的比例关系, 能更好地反映我国工业的发展状况. 本文选择 2010 年 1 月至 2021 年 2 月的月度工业增加率数据作为宏观经济变量之一.

两个宏观经济数据来源于《中国统计年鉴》, 图 1 和图 2 是居民消费价格指数水平值的时序图与工业增加率水平值的时序图.

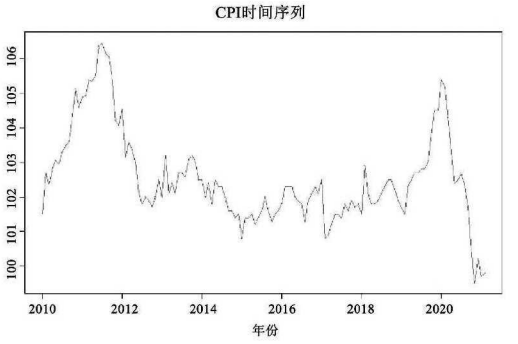


图 1 CPI 时序图

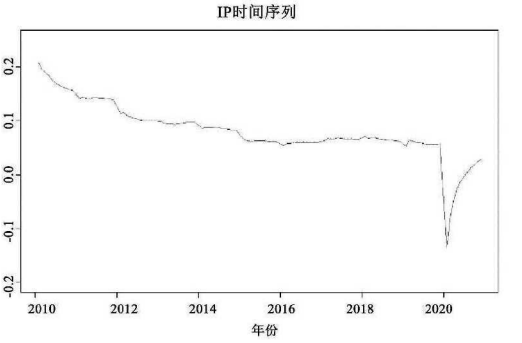


图 2 IP 时序图

根据图 1 得知, 对居民价格指数而言, 2011 年末到 2012 年初, 有一个大幅度的下跌, 在 2013 年后, 居民消费价格指数也在持续下跌; 2015 年到 2017 年间, 居民消费价格指数在震荡向上, 2017 年左右, 出现了断崖式的下跌, 跌到底部后出现反弹, 在 2018 年左右出现剧烈的上涨, 在此后, 居民消费价格指数持续上涨, 在 2020 年到达峰值后开始下跌. 由图 2 可知, 我国工业增加率是几乎平稳的时间序列, 且几乎都大于 0, 说明我国工业一直处于稳步上涨状态, 在 2020 年, 因新冠疫情, 我国工业生产受到了较大的影响, 出现了大幅度的下跌, 但是由于国家调控到位, 工业增加率迅速出现了反弹.

4.2 股市波动变量的选取

我国拥有上海证券交易所和深圳证券交易所两大证券交易所, 为了能更好反映我国股票市场的规模以及状况, 本文选择了上海证券交易所中最具市场代表性的 180 支股票编制的上证 180 指数 (简写 SSE180) 作为样本数据之一. 该指数能很好地反映上海证券交易所的运行状况. 同时为了反映另一个交易所的状况, 本文选取中证 100 指数 (简写 CSI100) 作为样本数据之一, 该指数反映了沪深两个交易所 100 支规模最大的股票的整体状况. 两个样本数据来源于万得金融数据库, 数据的时间区间选择 2010 年 1 月 4 日至 2021 年 2 月 26 日时间段,

中国知网 https://www.cnki.net

样本的频率均为日数据. 图 3 和图 4 分别是上证 180 收盘价的水平值和收益率的波动图. 图 5 和图 6 分别是中证 100 收盘价的水平值和收益率的波动图.

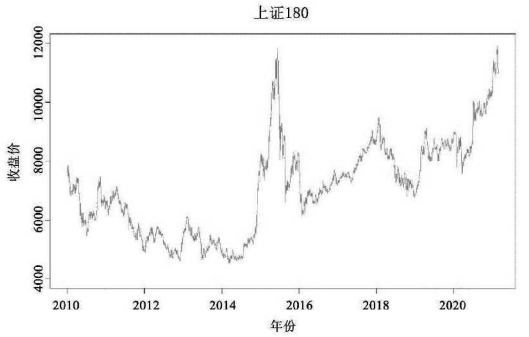


图 3 上证 180 的收盘价时序图

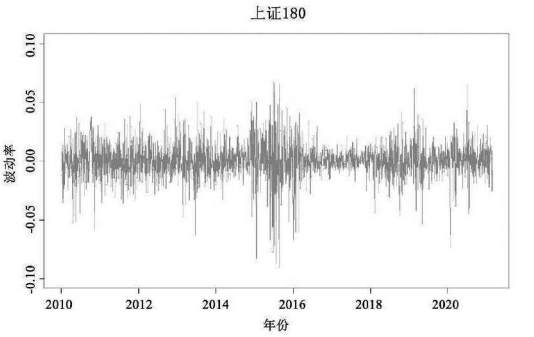


图 4 上证 180 收益率的波动图

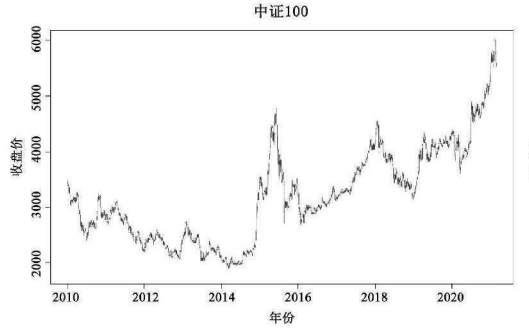


图 5 中证 100 的收盘价时序图

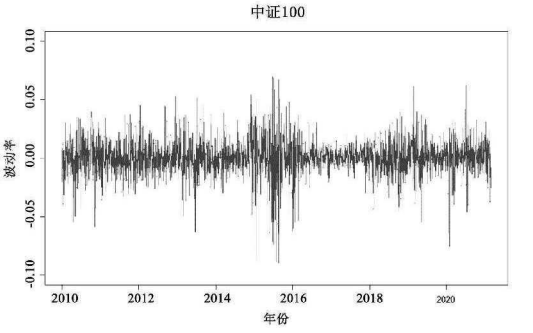


图 6 中证 100 收益率的波动图

由图 3 以及图 5 可知, 中国股市的波动性较大, 中国股票市场的剧烈波动集中在 2015 年和 2020 年附近, 在 2014 年至 2015 年间有一次股市的大涨, 以及 2020 年后股市也开始上涨. 观察图 4 以及图 6 的收益率的波动图可得知, 股市收益率的整体趋势较为平缓, 但其波动呈现出了聚集性, 较大的波动发生在同一个时间区间. 从整体上来看, 股市收益率序列是平稳的时间序列, 因此可以进一步对其波动率进行分析.

2014 年, 我国全面启动了深化体制改革, 改革为股市带来了重大的利好, 我国的股市进入了牛市, 但是在 2015 年时, 由于我国政策和资源配置不当, 股市由牛转熊, 我国为了稳定股票市场, 在 2016 年出台的挽救性措施是推行股市熔断制度, 但是效果却不尽如人意. 在制度推出后, 短期内 A 股便 “熔断”2 次, 市场恐慌情绪导致了股票价格的持续性下跌. 2019 年末, 新型冠状病毒疫情爆发, 为了拯救经济, 各国高层出台刺激经济发展政策, 并且随着去全球化市场的变化, 全球的经济政策较以前都变得更加宽松, 因此股市再次引来了上涨.

4.3 数据的波动处理

利用 GARCH-MIDAS 模型研究宏观经济因素的波动对股票市场波动影响前, 需要先对宏观经济数据进行预处理, 提取其波动率的值. 本文通过自回归模型来对宏观经济数据进行处理, 根据 AIC 和 BIC 信息最小的准则, 得到最优滞后阶数 p 建立 AR(p) 模型, 并对残差项进行白噪声检验, 提取出符合要求的残差项, 取其平方作为宏观经济因素时间序列的波动率,

中国知网 https://www.cnki.net

此时, 基于宏观经济数据波动率构建的 GARCH-MIDAS 方程为

$$\log \tau_t = m_v + \theta_v \sum_{k=1}^{K_v} \varphi_k(\omega_{1,v}, \omega_{2,v}) X_{v,t-k}^{mv}, \tag{8}$$

式中, $X_{v,t-k}^{mv}$ 代表滞后 k 期的宏观经济变量 X 的波动率.

图 7 和图 8 为居民消费价格指数和工业增加率两个宏观经济数据经过 AR(p) 模型滞后取残差的平方作为波动的波动率.

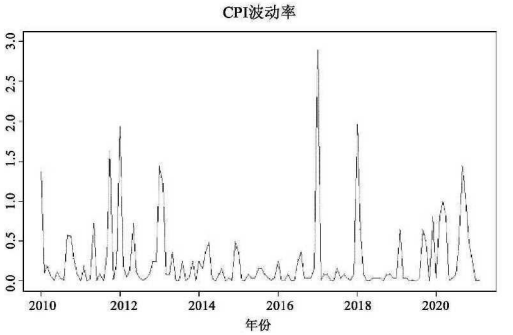


图 7　CPI 波动率图

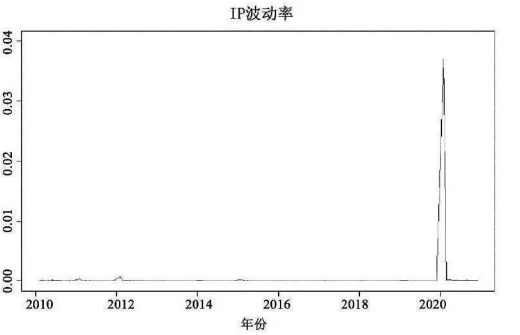


图 8　IP 波动率图

从图 7 和图 8 中可以看到, 居民消费价格指数长期处于波动的状态, 其波动较大的时间段在 2012 年, 2017 年, 2018 年以及 2020 年的附近. 我国工业增加率波动很小, 长期处于较为平稳的状态, 唯一出现极端波动的时间段为 2020 年年初.

4.4　数据描述性统计分析

表 1 给出了选取的四个变量的描述性统计的结果, 选取的统计量包括均值、最大值、最小值、标准差、变异系数、偏度、峰度和 Shapiro-Wilk 统计量, 具体结果如表 1 所示.

表 1　描述性统计分析

	CPI	IP	SSE180	CSI100
均值	102.55	0.068	7110.19	3123.844
最大值	106.45	0.207	11741.07	6030.517
最小值	99.5	–0.135	4545.14	1903.48
标准差	1.352	0.0508	1588.123	828.509
变异系数	0.013	0.747	0.223	0.265
偏度	0.899	–0.567	0.369	0.560
峰度	3.8249	6.0283	2.5572	2.8203
Shapiro-Wilk 统计量	0.918***	0.901***	0.964***	0.960***

注: *** 为 1% 水平下显著

表 1 中的标准差, CPI、IP、SSE180 和 CSI100 的标准差分别为 1.3526、0.0508、1588.123、828.5097, 且 Shapiro-Wilk 统计量均显著, 其 p 值均小于 0.01, 拒绝数据服从正态分布的原假设, 说明本文选取的两个宏观经济变量以及股票价格数据均不服从正态分布, 聚集效应不明显. 从变异系数来看, 变异系数越小, 说明变异程度越小, 其波动也越小, 变异系数越大, 数据

中国知网　https://www.cnki.net

中越可能出现离群值. 四个数据中, 工业增加率较大概率出现了离群值, 居民消费价格指数的波动最小, 两个股票价格数据波动程度相当. 居民消费价格指数以及股票价格数据的偏度均大于 0 且严重右偏, 而工业增加率的偏度小于 0, 严重左偏. 所选的四个数据峰度均为正数, 呈现长尾的特征, 具有明显的 "尖峰厚尾" 特征, 适合使用 GARCH-MIDAS 进行拟合.

5 波动性及动态影响因素分析

利用 GARCH-MIDAS 模型及上述模型参数的估计方法 (即: 极大似然估计法), 将股票市场的总波动率用长期波动来代替, 若采用固定窗口进行估计, 则长期成分固定为一个周期, 若采用滑动窗口进行估计则长期成分在每一天都会发生变化. 图 9 与图 10 分别为固定窗口下与滑动窗口下上证 180 指数波动率的总波动率与长期成分的情况, 图 11 与图 12 分别为固定窗口下与滑动窗口下中证 100 指数波动率的总波动率与长期成分的情况.

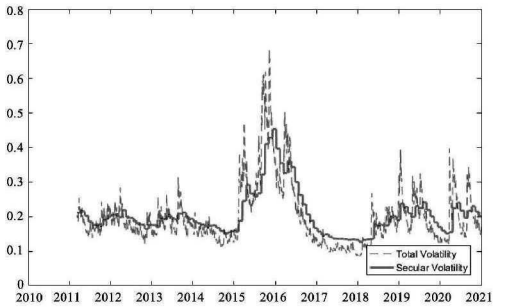


图 9 上证 180 波动率及长期成分 (固定窗口)

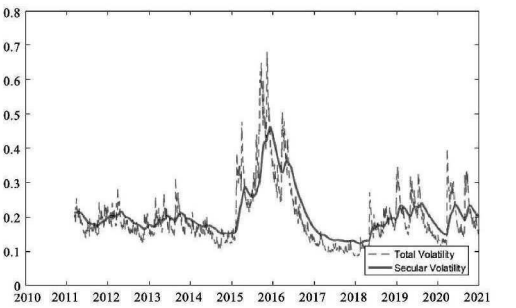


图 10 上证 180 波动率及长期成分 (滑动窗口)

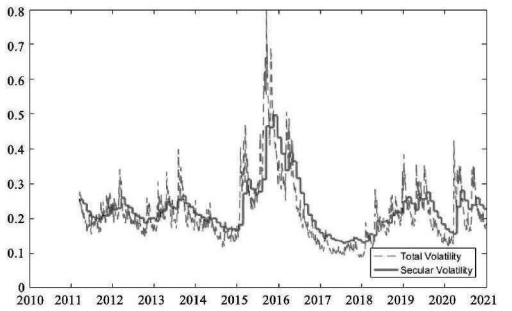


图 11 中证 100 波动率及长期成分 (固定窗口)

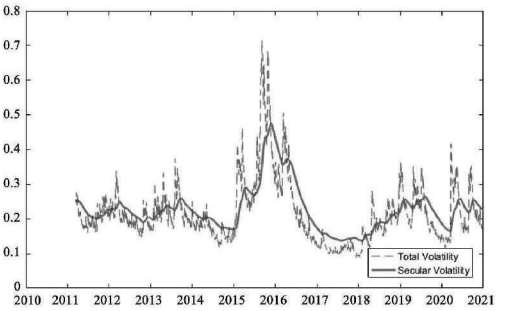


图 12 中证 100 波动率及长期成分 (滑动窗口)

由四幅图片可知, 无论是固定窗口进行估计, 还是用滑动的窗口进行估计, 股票市场的长期波动和其总波动率的趋势基本都是一致的, 这说明在一个短周期内, 保持常数不变, 或者让他在每天都变化, 对整体趋势的变化并无太大的影响. 由图也能看出, 长期波动较股票总波动更为平缓, 这是因为长期波动排除了部分噪声的干扰, 因此长期波动能更恰当地去描述股市的波动情况. 通过将波动率的长期成分代替总波动进行估计, 能更好地分析宏观经济因素对股市波动的长期影响.

将宏观经济数据作为解释变量, 根据宏观经济变量波动率 $X_{v,t-k}^{mv}$ 构建 GARCH-MIDAS 模型, 模型的估计结果如表 2, 表 3 所示.

中国知网 https://www.cnki.net

表 2 上证 180 的 GARCH-MIDAS 模型波动效应结果

变量	CPI	CPI (滑动)	IP	IP (滑动)
μ	0.000403*	0.000391*	0.000429**	0.000421**
	(1.916)	(1.8634)	(2.0218)	(1.9809)
α	0.068548***	0.061892***	0.07072***	0.07095***
	(16.8)	(14.884)	(16.209)	(16.331)
β	0.92889***	0.92757***	0.92335***	0.9234***
	(227.18)	(224.79)	(215.27)	(216.12)
θ	7.9334**	16.305***	3.362***	3.122***
	(1.8084)	(3.7283)	(3.1714)	(2.9122)
ω	45.846	49.559*	1.0011***	1.0439***
	(1.4006)	(1.7341)	(56.817)	(2.9849)
m	0.000043***	0.000033***	0.000034	0.000061
	(3.9978)	(4.0992)	(0.65001)	(0.91294)
AIC	–14719	–14730.9	–15317.4	–15315.1
BIC	–14683.6	–14695.5	–15282.1	–15279.7
LL	7365.5	7371.45	7664.71	7663.53
RMSE	0.00052	0.00052	0.00052	0.00052

注: * 为 10% 水平下显著, ** 为 5% 水平下显著, *** 为 1% 水平下显著, 括号内为对应的 T 值

表 3 中证 100 的 GARCH-MIDAS 模型波动效应结果

变量	CPI	CPI (滑动)	IP	IP (滑动)
μ	0.000521**	0.000517**	0.000533**	0.000534**
	(2.3722)	(2.3548)	(2.3891)	(2.4018)
α	0.069227***	0.063253***	0.07164***	0.07195***
	(16.821)	(14.907)	(15.76)	(16.595)
β	0.92759***	0.9333***	0.92226***	0.92148***
	(223.97)	(218.94)	(206.56)	(208.3)
θ	9.3315**	15.884***	2.6056***	3.2111***
	(2.128)	(3.8582)	(2.9334)	(3.1596)
ω	36.665*	49.889*	1.1573*	1.327*
	(0.14092)	(1.6517)	(1.2129)	(1.6937)
m	0.000399***	0.00032***	0.000098	0.00004
	(4.1982)	(4.3086)	(1.1818)	(0.5277)
AIC	–14776.1	–14786.6	–14755	–14776.4
BIC	–14740.7	–14751.1	–14739.6	–14741
LL	7394.07	7399.29	7393.49	7394.21
RMSE	0.000515	0.000513	0.000512	0.000512

注: * 为 10% 水平下显著, ** 为 5% 水平下显著, *** 为 1% 水平下显著, 括号内为对应的 T 值

中国知网 https://www.cnki.net

5.1 居民消费价格指数波动效应

表格中第二列与第三列为分析 CPI 波动对股市的影响, α 与 β 均显著小于 1, 且 $\alpha+\beta$ 接近 1, 根据 CPI 建立的模型对股市价格波动长期成分的拟合效果优秀. θ 值是最重要的参数, 表示选取的宏观经济变量对股市价格波动长期成分的影响, 从表中可得知, 无论是固定窗口还是滑动窗口, 模型估计的 θ 值均为大于 0, 说明 CPI 与波动的长期成分是正相关的, 即影响是正向的. CPI 代表着国内物价的水平, 假设物价持续波动, 在通货膨胀严重的背景下, 投资者为了保值, 将会把资金投入股市之中, 因此股价势必也会随之波动. 模型的 ω 参数为衰减速率, ω 越大, 说明衰减速率越快, ω 越小, 则衰减速率也越小. 模型估计的 ω 均显著大于 2, 表明宏观经济因素滞后项权重系数随着滞后时间的增加, 对股市波动的影响越来越小, 但距离当前时间越近, 宏观经济因素对股市的影响会越大. 说明了股市的波动对 CPI 发生的变化是十分敏感的, 短期内 CPI 的剧烈波动, 股市也会有强烈的反应, 具有较为明确的经济意义.

5.2 工业增加率波动效应

表格中第四列与第五列为分析 IP 波动对股市的影响, α 与 β 均显著小于 1, 且 $\alpha+\beta$ 接近 1, 根据 IP 建立的模型对股市价格波动长期成分的拟合效果优秀. 从模型估计结果上来看, 和 CPI 类似, 模型估计的 θ 为正数, 说明 IP 波动的对股市的影响也是正向关联性的. 工业增加率是衡量我国经济增长状况的重要指标, 衡量了我国工业制造生产的基本状况, 工业增加率在增长, 说明我国经济在发展, 国民收入也在不断提高, 这将会不断刺激投资者进行多元化的投资, 将会有相当一部分的资金流入股市之中. 工业增加率不断提高, 也说明我国的各种上市公司经营状况良好, 这也同样会吸引更多的投资者将资金投入股市之中, 从而造成股市价格的波动.

与 CPI 的模型估计结果相比, IP 的模型的 θ 更小, 说明 IP 对股市的波动效应冲击要小于 CPI, 相比于 CPI, 股票市场受工业增加率的影响更小. 从模型的权重函数估计结果来看, 工业增加率波动率对股市的短期影响程度不明显, 但是随着滞后期增加, 工业增加率的影响会变大. 工业生产是一个国家经济发展的基础, 若国家经济长期持续下行, 各个上市公司经营状况长期不佳, 投资者的信心必然会减少, 符合客观经济规律.

6 结 论

本文通过 GARCH-MIDAS 模型研究宏观经济因素对股票市场波动的影响, 首次选取工业增加率这一基础宏观经济指标进行实证研究, 研究结果表明选取的居民消费价格指数与工业增加率两个宏观经济指标都对股市波动具有显著影响, 对股市冲击影响明显, 且选取的指标对股市波动影响都是正向的. 在两个宏观经济因素中, 居民消费价格指数波动对股市的影响更为显著. 对股市长短期影响方面, 两个宏观经济因素各不相同, 居民消费价格指数波动对股市的短期影响更大, 而工业增加率波动会对股市产生长期的影响. 从波动效应分析可以看出, 宏观经济因素对股市波动的影响十分显著, 类似于居民消费价格指数与工业增加率的宏观经济因素在一定程度上反映着未来股市的预期, 本文相关研究成果可为金融监管提供建议与依据, 从而尽可能地避免剧烈的市场波动.

中国知网 https://www.cnki.net

参考文献

[1] 顾海峰, 周亚勇. 中证 500 股指期货推出降低了中国股市波动吗?—— 来自 2007–2016 年沪深股市的证据 [J]. 统计与信息论坛, 2019, 34(2): 105-113.

[2] 刘庞庞. 期权市场对现货市场波动率影响的实证分析 —— 基于上证 50ETF 期权上市前后的比较 [J]. 统计与信息论坛, 2017, 32(10): 50-58.

[3] Asgharian H, Christiansen C, Hou A J. Effects of macroeconomic uncertainty upon the stock and bond markets[J]. CREATES Research Papers, 2015(13): 10-16.

[4] 彭伟. 我国上市商业银行股票日收益风险价值研究 —— 基于 AR、HAR 和 MIDAS 模型的分析 [J]. 金融监管研究, 2013(3): 89-103.

[5] 周德才, 贾青, 李梓玮. 基于我国货币政策不确定性的股市波动长短期成分测度研究 [J]. 金融发展研究, 2017(5): 25-32.

[6] 石强, 杨一文, 刘雅凯. 基于 GARCH-MIDAS 模型的宏观经济与股市波动关系 [J]. 计算机工程与应用, 2019, 55(15): 257-262.

[7] 夏婷, 闻岳春. 经济不确定性是股市波动的因子吗?—— 基于 GARCH-MIDAS 模型的分析 [J]. 中国管理科学, 2018, 26(12): 1-11.

[8] 尹力博, 李勍. 投资者关注对人民币汇率价差波动的影响研究 —— 基于 GARCH-MIDAS 模型 [J]. 管理科学, 2017, 30(5): 147-159.

[9] 王江涛, 黄立玮, 崔翔宇. 基于自适应权重的混频 GARCH 模型及其应用 [J]. 统计研究, 2021, 38(5): 97-108.

[10] 李丹. 基于 GARCH 族模型的分级基金波动性分析 [J]. 统计与决策, 2017(13): 160-163.

[11] Asgharian H, Christiansen C, Hou A J. Macro-finance determinants of the Long-Run Stock-Bond correlation: The DCC-MIDAS specification[R]. Working Papers, 2014, 14(3): 617-642.

[12] Ashwani, Sheera V P. Indian stock market volatility and economic fundamentals: MIDAS approach[J]. Indian Journal of Finance, 2018, 12(8): 7-21.

[13] Yu X, Huang Y. The impact of economic policy uncertainty on stock volatility: Evidence from GARCH-MIDAS approach[J]. Physica A: Statistical Mechanics and Its Applications, 2021(3): 125794.

[14] Wang L, Ma F, Liu J, et al. Forecasting stock price volatility: New evidence from the GARCH-MIDAS model[J]. International Journal of Forecasting, 2020, 36(2): 684-694.

[15] 张一锋, 雷立坤, 魏宇. 羊群效应的新测度指数及其对我国股市波动的预测作用研究 [J]. 系统工程理论与实践, 2020, 40(11): 2810-2824.

[16] Pan Z, Liu L. Forecasting stock, return volatility: A comparison between the roles of short-term and long-term leverage effects[J]. Physica A: Statistical Mechanics and Its Applications, 2018, 492(15): 168-180.

[17] Cheng A, Yip I. China's macroeconomic fundamentals on stock market volatility: Evidence from Shanghai and Hong Kong[J]. International Journal of Productivity and Quality Management, 2017, 20(2): 1750014. 1-1750014.57.

[18] Bera A K, Higgins M L. ARCH models: Properties, estimation and testing[J]. Journal of Economic Surveys, 1993, 7(4): 305-366.

[19] Engle R F, Ghysels E. Stock market volatility and macroeconomic fundamentals[J]. Review of Economics & Statistics, 2013, 95(3): 776-797.

[20] 郑挺国, 尚玉皇. 基于宏观基本面的股市波动度量与预测 [J]. 世界经济, 2014, 37(12): 118-139.

中国知网 https://www.cnki.net

Consumer Price Index, Industrial Growth Rate and Stock Index Volatility Correlation Analysis Based on GARCH-MIDAS Model

XIE Xian-fen[1], HE Yi-chen[2], MAO Yi-jun[3], GU Wan-rong[3],
ZHANG Zi-ye[2], HUANG Tian-rui[4]

(1. School of Economics, Jinan University, Guangzhou 510632, China)
(2. School of Computer Science and Engineer, South China University of Technology, Guangzhou 510641, China)
(3. School of Mathematics and Information, South China Agricultural University, Guangzhou 510642, China)
(4. School of Computer, Guangdong University of Education, Guangzhou 510303, China)

Abstract: To explore the relationship between macroeconomy factors and stock market volatility, this paper selected the SSE 180 Index and the CSI 100 Index as the research objects and took the Consumer Price Index and the Industrial Growth Rate as the research variables, then used GARCH-MIDAS model to study the influence of macroeconomy on the long-run and short-run components of stock market volatility. The empirical study shows that the model can fit the stock market volatility well and can better reflect the correlation between macroeconomic variables and stock market volatility. Both variables have a significant positive impact on the long-run components of stock volatility. The two have different effects on short-run components of stock market volatility. Industrial Growth Rate has no obvious impact on short-run components of stock market volatility, while short-run components of stock market volatility are very sensitive to the fluctuation of Consumer Price Index.

Keywords: stock market volatility; GARCH-MIDAS; mixed frequency data; long and short-run volatility

中国知网 https://www.cnki.net

SCAULIB202518174

检索证明

根据委托人提供的论文材料，委托人**华南农业大学数学与信息学院 古万荣** 1 篇论文收录情况如下表。

序号	论文名称	发表刊物及发表的年月卷期/页码等	作者排名	论文等级	作者文中单位	收录情况	影响因子	中科院大类分区
1	居民消费价格指数、工业增加率与股票指数波动关系分析——基于 GARCH-MIDAS 模型	数学的实践与认识 出版年：2025 出版日期：2025-03-12 卷期：55 3 页码：1-12 文献号： 文献类型：	通讯作者	C 类	华南农业大学	北大核心	无	无

说明：论文等级和中科院大类分区按《华南农业大学学术论文评价方案（试行）》划分。

报告免责声明:如未盖章,报告无效

检索员：田成

华南农业大学图书馆

2025-06-11

华南农业大学图书馆SCAULIB202518174

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号： 软著登字第7573092号

软 件 名 称： 面向X光图像的肺炎识别和定位系统
[简称：X光肺炎预测及定位系统]
V1.0

著 作 权 人： 华南农业大学

开发完成日期： 2020年04月15日

首次发表日期： 未发表

权利取得方式： 原始取得

权 利 范 围： 全部权利

登 记 号： 2021SR0850466

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局
计算机软件著作权登记专用章

No. 08150217

2021年06月08日

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号： 软著登字第7581695号

软 件 名 称： 果实识别系统
V1.0

著 作 权 人： 华南农业大学

开发完成日期： 2020年04月20日

首次发表日期： 2020年05月01日

权利取得方式： 原始取得

权 利 范 围： 全部权利

登 记 号： 2021SR0859069

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局
计算机软件著作权登记专用章

No. 08308871

2021年06月09日

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号：软著登字第2853820号

软件名称：基于faster r-cnn的甲状腺肿瘤超声图像自动识别系统
[简称：甲状腺超声图像肿瘤自动识别系统]
V1.0

著作权人：华南农业大学

开发完成日期：2018年03月01日

首次发表日期：2018年04月01日

权利取得方式：原始取得

权利范围：全部权利

登记号：2018SR524725

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局
计算机软件著作权
登记专用章

No. 02767385

2018年07月06日

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号：软著登字第2854890号

软 件 名 称：基于Nutch的互联网信息抓取系统
[简称：互联网信息抓取系统]
V1.0

著 作 权 人：华南农业大学

开发完成日期：2018年03月01日

首次发表日期：2018年04月01日

权利取得方式：原始取得

权 利 范 围：全部权利

登 记 号：2018SR525795

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局
计算机软件著作权
登记专用章
2018年07月06日

No. 02767002

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号：软著登字第2997234号

软 件 名 称：易书网系统
[简称：易书网]
V2.1

著 作 权 人：华南农业大学

开发完成日期：2018年03月01日

首次发表日期：2018年04月01日

权利取得方式：原始取得

权 利 范 围：全部权利

登 记 号：2018SR668139

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

No. 02941394

中华人民共和国国家版权局
计算机软件著作权
登记专用章
2018年08月21日

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号：软著登字第4287777号

软件名称：“百租易”物品租赁平台软件
[简称：百租易/Buyless]
V1.0

著作权人：华南农业大学

开发完成日期：2019年03月01日

首次发表日期：未发表

权利取得方式：原始取得

权利范围：全部权利

登记号：2019SR0867020

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局
计算机软件著作权
登记专用章

No. 04408652

2019年08月21日

华南农业大学 South China Agricultural University

技术成果申请专利（软件著作权）审批

拟申请软件著作权名　　称	“百租易”物品租赁平台软件		
申请人	华南农业大学	申请类型	软件著作权
联 系 人	古万荣	手机号码	13719398640
E-Mail	40840582@qq.com		
发明人			
第一发明人	古万荣	性　　别	男
注：学校给予本校教职工为第一发明人的知识产权配套奖励；			
所在单位	数学与信息学院、软件学院	职　　称	讲师

姓　　名	性　　别	所在单位	职　　称
古万荣	男	数学与信息学院、软件学院	讲师
吕浩泰	男	数学与信息学院、软件学院	本科生
毛宜军	男	数学与信息学院、软件学院	讲师
王国华	男	数学与信息学院、软件学院	讲师
梁早清	男	数学与信息学院、软件学院	讲师

支持课题情况				
课题名称	协同超图推荐模型研究及应用			
课题来源及编号	广东省自然科学基金项目-E18037-2018A030313437			
申请类别	☐发明专利　☐实用新型专利　☐外观设计专利			
代理机构名称	广州市华学知识产权代理有限公司中山分公司 代理机构资质查询 ： http://www.acpaa.cn/view/agencySearch.jhtml			
拟申请软件著作权简　　介	本软件实现了基于定位、移动平台的租赁推荐软件，通过智能匹配算法，用户可以将闲置物品放在租赁平台出租、出售，系统自动进行二分匹配计算并给用户推荐最佳的物品集合。			
发明人承诺	本人将按照学校有关规定，积极配合学校知识产权办的工作，做好专利（软件著作权）申请过程中的答复等相关事宜，并及时将收到的有关通知、材料交学校知识产权办处理。			
	是否为第一发明人	○是　○否	日　　期	2019-05-14

所在单位审核	本人代表学院对该专利（软件著作权）申请进行了审查。 学院将积极配合学校督促发明人按照学校的有关规定和工作需要做好该专利（软件著作权）的申请、宣传、转化等工作。			
		审核人	黄琼	审核时间	2019-05-14

科技处业务科室科员审核	科技处已备存，无需提交纸质版专利（软件著作权）审批表；如有材料需加盖公章，请在办事服务大厅--知识产权类材料用章申请办理			
		审核人	刘毅新	审核时间	2019-05-14

您（古万荣）填写/办理了申请人确认（完成）　2019/05/15 16:59

☐ 点击展开5个较早历史

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号：软著登字第4290992号

软件名称：智能问答系统软件
[简称：智能问答系统]
V1.0

著作权人：华南农业大学

开发完成日期：2019年04月25日

首次发表日期：未发表

权利取得方式：原始取得

权利范围：全部权利

登记号：2019SR0870235

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局
计算机软件著作权
登记专用章
2019年08月21日

No. 04397748

華南農業大學
South China Agricultural University

技术成果申请专利（软件著作权）审批

拟申请软件著作权名　　称	智能问答系统软件		
申请人	华南农业大学	申请类型	软件著作权
联 系 人	古万荣	手机号码	13719398640
E-Mail	40840582@qq.com		
发明人			
第一发明人	古万荣	性　　别	男
注：学校给予本校教职工为第一发明人的知识产权配套奖励；			
所在单位	数学与信息学院、软件学院	职　　称	讲师

姓　　名	性　　别	所在单位	职　　称
古万荣	男	数学与信息学院、软件学院	讲师
李敏诗	女	数学与信息学院、软件学院	本科生
何亦琛	男	数学与信息学院、软件学院	本科生
何岱俊	男	数学与信息学院、软件学院	本科生
毛宜军	男	数学与信息学院、软件学院	讲师

支持课题情况				
课题名称	协同超图推荐模型研究及应用			
课题来源及编号	广东省自然科学基金项目E18037-2018A030313437			
申请类别	☐发明专利　☐实用新型专利　☐外观设计专利			
代理机构名称	广州市华学知识产权代理有限公司中山分公司 代理机构资质查询 ： http://www.acpaa.cn/view/agencySearch.jhtml			
拟申请软件著作权简　　介	本软件实现了一个智能问答系统，可经过知识库的整理，回答面向特定领域的问答，也可以通过互联网搜索，将答案进行整理，回答通用问题。			
发明人承诺	本人将按照学校有关规定，积极配合学校知识产权办的工作，做好专利（软件著作权）申请过程中的答复等相关事宜，并及时将收到的有关通知、材料交学校知识产权办处理。			
	是否为第一发明人	○是　○否	日　　期	2019-05-14

所在单位审核	本人代表学院对该专利（软件著作权）申请进行了审查。 学院将积极配合学校督促发明人按照学校的有关规定和工作需要做好该专利（软件著作权）的申请、宣传、转化等工作。			
		审核人	黄琼	审核时间	2019-05-14

科技处业务科室科员审核	科技处已备存，无需提交纸质版专利（软件著作权）审批表；如有材料需加盖公章，请在办事服务大厅--知识产权类材料用章申请办理			
		审核人	刘毅新	审核时间	2019-05-14

您（古万荣）填写/办理了申请人确认（完成）　2019/05/15 17:00

☐ 点击展开5个较早历史

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号：软著登字第4290982号

软件名称：个人事务管理系统
V1.0

著作权人：华南农业大学

开发完成日期：2019年04月25日

首次发表日期：未发表

权利取得方式：原始取得

权利范围：全部权利

登记号：2019SR0870225

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局
计算机软件著作权
登记专用章
2019年08月21日

No. 04397747

華南農業大學 South China Agricultural University

技术成果申请专利（软件著作权）审批

拟申请软件著作权名　　称	个人事务管理系统		
申请人	华南农业大学	申请类型	软件著作权
联 系 人	古万荣	手机号码	13719398640
E-Mail	40840582@qq.com		

发明人

第一发明人	古万荣	性　别	男

注：学校给予本校教职工为第一发明人的知识产权配套奖励；

所在单位	数学与信息学院（软件学院）	职　称	讲师

姓　　名	性　别	所在单位	职　　称
古万荣	男	数学与信息学院（软件学院）	讲师
张怡	女	数学与信息学院（软件学院）	本科生
何亦琛	男	数学与信息学院（软件学院）	本科生
李吉平	男	数学与信息学院（软件学院）	教授

支持课题情况

课题名称	基于复杂对象关系模型的新闻推荐研究
课题来源及编号	广东省社科项目-E19014（GD18CXW01）
申请类别	☐发明专利　☐实用新型专利　☐外观设计专利
代理机构名称	广州市华学知识产权代理有限公司中山分公司 代理机构资质查询：http://www.acpaa.cn/view/agencySearch.jhtml
拟申请软件著作权简　　介	本软件实现了基于微信小程序的个人事务管理软件，该软件除了实现了基本的事务管理功能外，还结合网络信息和用户偏好，为社交、新闻推荐、复杂社交模型打下数据基础。
发明人承诺	本人将按照学校有关规定，积极配合学校知识产权办的工作，做好专利（软件著作权）申请过程中的答复等相关事宜，并及时将收到的有关通知、材料交学校知识产权办处理。

是否为第一发明人	◉是　○否	日　　期	2019-05-14

所在单位审核	本人代表学院对该专利（软件著作权）申请进行了审查。 学院将积极配合学校督促发明人按照学校的有关规定和工作需要做好该专利（软件著作权）的申请、宣传、转化等工作。			
		审核人	黄琼	审核时间	2019-05-14

科技处业务科室科员审核	科技处已备存，无需提交纸质版专利（软件著作权）审批表；如有材料需加盖公章，请在办事服务大厅--知识产权类材料用章申请办理			
		审核人	刘毅新	审核时间	2019-05-14

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号：软著登字第4291000号

软件名称：HTML旅游网站软件
[简称：旅游网站]
V1.0

著作权人：华南农业大学

开发完成日期：2019年04月26日

首次发表日期：未发表

权利取得方式：原始取得

权利范围：全部权利

登记号：2019SR0870243

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局
计算机软件著作权
登记专用章
2019年08月21日

No. 04397749

华南农业大学
South China Agricultural University

技术成果申请专利（软件著作权）审批

拟申请软件著作权名　　称	HTML旅游网站		
申请人	华南农业大学	申请类型	软件著作权
联 系 人	古万荣	手机号码	13719398640
E-Mail	40840582@qq.com		
发明人			
第一发明人	古万荣	性　　别	男
注：学校给予本校教职工为第一发明人的知识产权配套奖励；			
所在单位	数学与信息学院、软件学院	职　　称	讲师

姓　　名	性　　别	所在单位	职　　称
古万荣	男	数学与信息学院、软件学院	讲师
陈卓均	男	数学与信息学院、软件学院	本科生
黄诗诗	女	数学与信息学院、软件学院	本科生
何亦琛	男	数学与信息学院、软件学院	本科生
梁早清	男	数学与信息学院、软件学院	讲师
毛宜军	男	数学与信息学院、软件学院	讲师
王国华	男	数学与信息学院、软件学院	讲师

支持课题情况				
课题名称	协同超图推荐模型研究及应用			
课题来源及编号	广东省自然科学基金项目E18037-2018A030313437			
申请类别	☐发明专利　☐实用新型专利　☐外观设计专利			
代理机构名称	广州市华学知识产权代理有限公司中山分公司 代理机构资质查询 ： http://www.acpaa.cn/view/agencySearch.jhtml			
拟申请软件著作权简　　介	本软件设计和实现了一个旅游管理网站，主要功能涉及到酒店管理、线路管理、游客信息管理等。对于独立景区及其景区的一体化管理具有实用性。			
发明人承诺	本人将按照学校有关规定，积极配合学校知识产权办的工作，做好专利（软件著作权）申请过程中的答复等相关事宜，并及时将收到的有关通知、材料交学校知识产权办处理。			
	是否为第一发明人	○是　○否	日　　期	2019-05-14

所在单位审核	本人代表学院对该专利（软件著作权）申请进行了审查。 学院将积极配合学校督促发明人按照学校的有关规定和工作需要做好该专利（软件著作权）的申请、宣传、转化等工作。			
		审核人	黄琼	审核时间	2019-05-14

科技处业务科室科员审核	科技处已备存，无需提交纸质版专利（软件著作权）审批表；如有材料需加盖公章，请在办事服务大厅--知识产权类材料用章申请办理			
		审核人	刘毅新	审核时间	2019-05-14

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号：软著登字第4290976号

软件名称：基于iOS平台的图像社交软件
[简称：图像社交软件]
V1.0

著作权人：华南农业大学

开发完成日期：2019年04月10日

首次发表日期：未发表

权利取得方式：原始取得

权利范围：全部权利

登记号：2019SR0870219

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局
计算机软件著作权
登记专用章
2019年08月21日

No. 04397746

华南农业大学 South China Agricultural University

技术成果申请专利（软件著作权）审批

拟申请软件著作权名　　称	基于iOS平台的图像社交软件		
申请人	华南农业大学	申请类型	软件著作权
联 系 人	古万荣	手机号码	13719398640
E-Mail	40840582@qq.com		

发明人

第一发明人	古万荣	性　别	男

注：学校给予本校教职工为第一发明人的知识产权配套奖励；

所在单位	数学与信息学院（软件学院）	职　称	讲师

姓　　名	性　别	所在单位	职　　称
古万荣	男	数学与信息学院（软件学院）	讲师
王玉亭	女	数学与信息学院（软件学院）	本科生
黄诗诗	女	数学与信息学院（软件学院）	本科生
毛宜军	男	数学与信息学院（软件学院）	讲师
梁早清	男	数学与信息学院（软件学院）	讲师

支持课题情况

课题名称	基于超图模型的新闻推荐研究及应用研究
课题来源及编号	教育部人文社科项目-C18087（18YJCZH037）
申请类别	☐ 发明专利　☐ 实用新型专利　☐ 外观设计专利
代理机构名称	广州市华学知识产权代理有限公司中山分公司 代理机构资质查询：http://www.acpaa.cn/view/agencySearch.jhtml
拟申请软件著作权简　　介	本软件实现了一个基于ios系统平台的app，该软件可以实现基本的社交功能，满足小众群体的社交、沟通存储、文件传输等基本功能。
发明人承诺	本人将按照学校有关规定，积极配合学校知识产权办的工作，做好专利（软件著作权）申请过程中的答复等相关事宜，并及时将收到的有关通知、材料交学校知识产权办处理。

是否为第一发明人	◉ 是　○ 否	日　　期	2019-05-14

所在单位审核	本人代表学院对该专利（软件著作权）申请进行了审查。 学院将积极配合学校督促发明人按照学校的有关规定和工作需要做好该专利（软件著作权）的申请、宣传、转化等工作。			
	审核人	黄琼	审核时间	2019-05-14

科技处业务科室科员审核	科技处已备存，无需提交纸质版专利（软件著作权）审批表；如有材料需加盖公章，请在办事服务大厅--知识产权类材料用章申请办理			
	审核人	刘毅新	审核时间	2019-05-14

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号：软著登字第2853551号

软件名称：基于词向量模型的查询词扩展搜索平台
[简称：查询词扩展搜索平台]
V1.0

著作权人：华南农业大学

开发完成日期：2018年03月01日

首次发表日期：2018年04月01日

权利取得方式：原始取得

权利范围：全部权利

登记号：2018SR524456

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局
计算机软件著作权登记专用章

No. 02767350

2018年07月06日

证书号第7353998号

专利公告信息

发明专利证书

发明名称：一种基于改进的神经网络实现食品风险评估的方法

专利权人：华南农业大学

地　　址：510642 广东省广州市天河区五山路483号

发 明 人：柯海萍;毛宜军;古万荣;梁早清;黄锦涛;李观明;陈蔚钊

专　利　号：ZL 2021 1 1013269.X　　授权公告号：CN 113887868 B

专利申请日：2021年08月31日　　授权公告日：2024年09月06日

申请日时申请人：华南农业大学

申请日时发明人：柯海萍;毛宜军;古万荣;梁早清;黄锦涛;李观明;陈蔚钊

国家知识产权局依照中华人民共和国专利法进行审查，决定授予专利权，并予以公告。

专利权自授权公告之日起生效。专利权有效性及专利权人变更等法律信息以专利登记簿记载为准。

局长
申长雨

国家知识产权局
2024年09月06日

蓝桥杯大赛

获奖证书

华南农业大学古万荣：

指导汪凯旋荣获第十四届蓝桥杯全国软件和信息技术专业人才大赛全国总决赛C/C++程序设计大学B组三等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：021454859

证件号码：440229198201270018

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2023年6月16日

蓝桥杯大赛

获奖证书

华南农业大学古万荣：

指导郑晓楊荣获第十四届蓝桥杯全国软件和信息技术专业人才大赛广东赛区C/C++程序设计大学A组一等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：011412594

证件号码：440229198201270018

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2023年4月23日

蓝桥杯大赛

获奖证书

华南农业大学古万荣：

指导汪凯旋荣获第十四届蓝桥杯全国软件和信息技术专业人才大赛广东赛区C/C++程序设计大学B组一等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：021435688

证件号码：440229198201270018

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2023年4月23日

蓝桥杯大赛

获奖证书

华南农业大学古万荣：

指导陈煜栋荣获第十四届蓝桥杯全国软件和信息技术专业人才大赛广东赛区C/C++程序设计大学B组二等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：021435737

证件号码：440229198201270018

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2023年4月23日

蓝桥杯大赛

获奖证书

华南农业大学古万荣：

指导胡煜荣获第十四届蓝桥杯全国软件和信息技术专业人才大赛广东赛区C/C++程序设计大学B组三等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：021435630

证件号码：440229198201270018

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2023年4月23日

蓝桥杯大赛

获奖证书

华南农业大学古万荣：

指导谭文锋荣获第十四届蓝桥杯全国软件和信息技术专业人才大赛广东赛区C/C++程序设计大学B组三等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：021435667

证件号码：440229198201270018

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2023年4月23日

蓝桥杯大赛

获奖证书

华南农业大学古万荣：

指导陈少游荣获第十四届蓝桥杯全国软件和信息技术专业人才大赛广东赛区C/C++程序设计大学B组三等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：021435748

证件号码：440229198201270018

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2023年4月23日

获奖证书

华南农业大学

查云起、陈国伟、郭 丹：

你（们）的作品《基于 RAG 的农业智能问答与文档生成系统》在第十九届“挑战杯”全国大学生课外学术科技作品竞赛“人工智能 +”专项赛挑战赛中荣获

三等奖

指导教师：古万荣、王建华

编号：2025-TZB19-RT30124H-9C46B1

特颁此证，以兹鼓励

主办单位：共青团中央、中国科协、教 育 部、中国社科院
中国工程院、全国学联、江苏省人民政府

第十九届“挑战杯”全国大学生
课外学术科技作品竞赛
全国组织委员会
二〇二五年十一月

荣誉证书

古万荣老师

在华南农业大学2018年暑期“三下乡”社会实践活动中表现突出，被评为 优秀指导老师 。

特发此证，以资鼓励。

共青团华南农业大学委员会

2018年11月

荣誉证书

古万荣 同志：

在2018年度 科研 工作中成绩突出，授予“十佳工作者”称号。特发此证，以资鼓励。

华南农业大学数学与信息学院

二〇一九年一月