计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2019.20180641
56(7): 1383-1395, 2018

一种基于集成学习的科研合作者潜力预测分类方法

艾 科 马国帅 杨凯凯 钱宇华
(山西大学大数据科学与产业研究院 太原 030006)
(计算智能与中文信息处理教育部重点实验室(山西大学) 太原 030006)
(山西大学计算机与信息技术学院 太原 030006)
(aike0229@163.com)

A Classification Method of Scientific Collaborator Potential Prediction Based on Ensemble Learning

Ai Ke, Ma Guoshuai, Yang Kaikai, and Qian Yuhua
(*Institute of Big Data Science and Industry, Shanxi University, Taiyuan* 030006)
(*Key Laboratory of Computational Intelligence and Chinese Information Processing (Shanxi University), Ministry of Education, Taiyuan* 030006)
(*School of Computer and Information Technology, Shanxi University, Taiyuan* 030006)

Abstract Scientific cooperation is a very important form of academic achievement. Many high-level researches are achieved through cooperation. Researching the collaboration potential can provide guidance for scholars to choose collaborators and maximize the efficiency of scientific research. However, the current outbursts of big data have hindered the effective choice of collaborators. In order to solve the problem, based on scholar-paper big data, after features analysis and optimization and comprehensively considering individual attributes and related attributes of scholars' papers, institutions, research interests, etc., sample features from various dimensions such as paper title, paper rank, paper number, time and coauthor order are constructed. Taking journal or conference level of papers as the sample tags of collaborators sequence pairs, which indicates the potential of current cooperators and make use of the strong learning characteristics of the ensemble methods, a scientific collaborator potential prediction model based on ensemble learning classification method is proposed. After analyzing and constructing the feature set that corresponds to the problem of scientific collaborator potential prediction, classification method is adopted to solve the problem. In experiments, the accuracy, recall rate, and F1 score are much higher than those of traditional machine learning methods and can converge to high values (above 80%) with few samples and little time, indicating the superiority of the proposed model.

Key words scientific cooperation; potential prediction; feature construction; big scholar data; ensemble learning

摘 要 科研合作是学术成果非常重要的实现形式,很多高水平的研究成果通过合作实现.研究合作潜力可以为学者选择合作者提供指导,最大化科研效率.然而当前大数据爆发阻碍了合作者的有效选择.为了解决这个问题,基于学者-文章大数据,经过特征分析和优化,综合考虑学者的文章、机构、研究兴趣

收稿日期:2018-09-12;**修回日期**:2018-11-19
基金项目:国家自然科学基金项目(61672332,61432011,U1435212);山西省海外归国人员研究项目(2017023)
This work was supported by the National Natural Science Foundation of China (61672332, 61432011, U1435212) and the Overseas Returnee Research Project of Shanxi Scholarship Council of China (2017023).
通信作者:钱宇华(jinchengqyh@sxu.edu.cn)

等个人属性和相关属性,分别从文章标题、文章等级、文章数量、时间及署名序多维度构造样本特征,以文章所发表的期刊会议等级作为合作者序列对的样本标签,表示当前合作者的潜力高低,利用集成方法的强学习特性,提出了基于集成学习分类方法的科研合作者潜力预测模型.分析并构造对应于科研合作者潜力预测问题的特征集后,采用分类方法解决这一问题.实验中准确率、召回率、F1 分数都远高于传统机器学习方法,并能以较少的样本和时间收敛于较高值(80%以上),说明了模型的优越性.

关键词 科研合作;潜力预测;特征构造;学术大数据;集成学习

中图法分类号 TP18; TP391

科学学[1](science of science)旨在发掘学科的发展动力,构造模型反映学科演化过程,进而推动科学事业发展,科研合作就是其研究内容之一.合作和产出之间有强相关性[2],越来越多的高水平研究成果通过合作实现,这正凸显了合作者选择的重要性.优秀的合作关系能够充分发挥各合作者的潜力,最大化科研效益.通过科研合作模式指导,预先甄别合作者的潜力有助于学者平衡投入与产出,选择潜在收益最大的合作者,最大化科研效率.

合作关系所形成的合作数据反映了学者间的相互关系,是科研网络和学者行为的重要研究对象.基于合作数据的科研合作模式研究是当前的热点内容.科研合作模式对于研究学者行为有着非常重要的意义.以科研合作模式为载体的合作者推荐问题研究,多基于复杂网络理论[3];以点和边的拓扑分析为基础,把合作者推荐问题作为链路预测问题[4]处理,如基于随机游走的最有价值合作者 MVCWalker (most valuable collaborator)[5]方法预测二者之间产生合作的可能性;Tang 等人[6]则致力于解决交叉学科的合作者推荐和预测问题;一些其他方法也得到了较好的效果,如把共同参加同一会议作为影响合作产生的因素[7],以统计概率的形式描述新合作的产生;以及通过量化学者间的局部相关性和全局相似性[8]进行合作者推荐;模式识别方面,Xia 等人[9]利用高维度多角度的学术大数据,结合数据特征构造 Shifu 模型,对导师-学生关系进行挖掘.

然而以上的工作多以合作产生的可能性为研究目的,并没有对合作的结果给出预判性指导.为了达到最好的合作效果,需要对学者的合作潜力进行研究.但是仅仅依靠传统的拓扑关系已经无法满足问题需求,需要质量更高、信息量更大的数据来支撑.然而学术大数据(big scholar data, BSD)[10]的"爆发"性质[11]使得合作者潜力预测问题成为挑战.首先,数据量巨大使模式挖掘更加困难;其次,数据形式多样,不局限于现有方法中使用的结构化数据,学术大数据包含许多异构信息,如作者、文章、机构、期刊会议等,以及合作者合著关系等复杂网络关系[12];同时,数据具有动态性,学者个人、文章的影响力以及学者之间的合作强度都是与时间相关的,时间不同效果也不同.来自问题和数据的多重挑战迫切需要提出更有效的解决方法.

集成学习算法[13]是机器学习的一种新学习思想,该学习算法把同一个问题分解到多个不同模块中,由多个学习器参与学习,共同解决目标问题,最终通过平均或投票选用分类器,从而提高分类器泛化能力.根据个体学习器的生成方式,目前的集成学习方法大致可分为两大类,即个体学习器间存在强依赖关系、必须串行生成的序列化方法(以 Boosting[14]为代表),以及个体学习器间不存在强依赖关系、可同时生成的并行化方法(Bagging[15]和随机森林[16](random forest, RF)为代表).

因此,本文把集成学习分类方法应用于真实学者-文章大数据,构造面向合作者潜力预测问题的样本集.样本特征综合考虑学者的个人属性以及合作者之间的相关性,分别从文章标题、文章等级、文章数量、时间、署名序等多维度进行特征构造[17-18],进而提出了基于集成学习分类方法的科研合作者潜力预测模型.该模型旨在通过学者的属性集,对当前合作者的潜力进行预测.

本文的主要贡献有 2 个方面:

1) 构造了面向合作者潜力预测模型的样本集.将学术大数据中的文章、作者与文章等级进行对应,处理成含等级的文章和含等级的作者数据作为基准数据集.同时定义了一系列学术背景下的学者个人特征描述以及学者间相关性特征描述.

2) 提出合作者潜力预测的挖掘模型.将分类方法应用于以上特征集来解决合作者的潜力预测问题且实验效果显著.

1 合作者潜力预测模型设计思路

本文提出的合作者潜力预测模型基于假设:合作者潜力通过合作成果的等级高低表现,而成果等级

高低与合作者各自的单一属性和合作者之间的相关属性密切相关——作为个体每个学者都有一系列学术属性和社会属性,一定存在某种潜在模式使合作者的属性合集达到特定的形式时会产出特定等级的合作成果,这正是模型构建的出发点.本文通过对真实学术大数据进行分析并构建样本,利用集成学习算法在样本集挖掘合作者潜力预测模型.模型整体流程如图 1 所示:

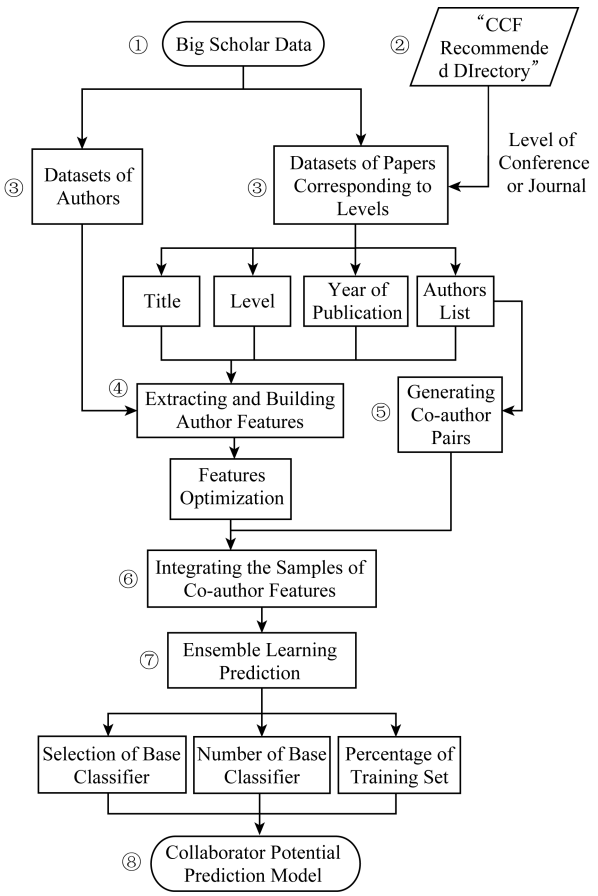


Fig. 1　Flow chart of the model

图 1　模型流程图

① 以 ArnetMiner[19] 提供的学术社会网络数据集为数据源提取特征,其中包括学者数据集和文章数据集.

② 以《中国计算机学会推荐国际学术会议和期刊目录》(简称《CCF 推荐目录》)中的类别 A,B,C(其中 A 代表高水平的期刊和会议)作为评价合作成果等级的标签.

③ 以作者为唯一标识符构建作者数据集,以文章为唯一标识符结合《CCF 推荐目录》构建包含等级的文章数据集.

④ 基于格式化的作者和文章数据集抽取和构建作者包含等级的作者特征集.考虑文章标题、等级、发表年份、作者列表属性,从时间、数量、文本相似性等角度综合度量不同特征对结果产生的偏差.

作者属性如题目、研究兴趣等都是文本形式,为了分析这些文本特征之间的关系,本文利用了自然语言处理中的潜在语义索引(latent semantic indexing, LSI)模型[20].LSI 基于奇异值分解(singular value decomposition, SVD)的方法得到文本主题,通过 1 次 SVD 过程得到文档和主题的相关度、词和词义的相关度以及词义和主题的相关度索引.

⑤ 科研合作是由合著关系表示的一种强社会关系.不同模式的合作关系隐藏在广泛的科研合作关系中.在共著关系基础上可以构建 1 个科研合作网络[21],如图 2 所示.在科研合作网络中,2 位学者如果共同撰写文章就会被认为是相互联系的.基于科研合作网络的结构和社会规律对合作模式挖掘非常重要.

类似于网络图中对边的定义和研究,模型构建中只考虑合作成果作为 1 条边的情况,即只考虑 2 个人而非多个人之间产生合作的模型构建.

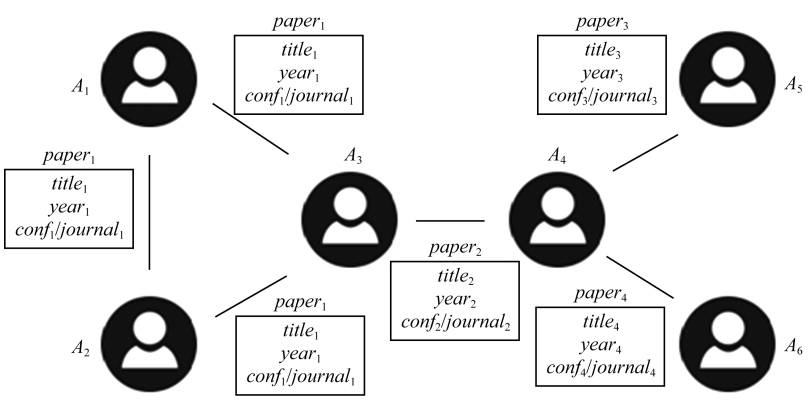


Fig. 2　Scientific coauthor network

图 2　科研合作网络

⑥ 构造基于当前合作双方的基准特征集,包含作者各自的特征以及二者之间相关性的特征.若合作样本(文章)的等级为 A,则样本标签记为"1",否则为"0",进而将特征和样本整合为样本集.

⑦ 分别采用 Boosting,Bagging,RF 这 3 种基分类器对以上样本集进行集成学习.分别改变训练集比例以及基分类器个数以测试所构造样本集在当前研究中的有效性.

⑧ 得到学术大数据下的科研合作者潜力预测模型.

2 合作者潜力预测模型构建过程

本文构建合作者潜力预测模型的过程主要包含 2 部分:1)基于科研合作大数据的分析,提取可用特征;2)基于学者基本数据构造特征样本,采用集成学习算法构建模型,完成合作者潜力预测的任务.

2.1 特征分析

考虑与合作者潜力相关的因素,以统计图表形式分别对 4 个特征进行分析:1)文章标题;2)不同等级中的文章数量;3)文章发表年份;4)署名序.

首先,图 3 以发表篇数为横坐标、发表该篇数的人数为纵坐标初步刻画了数据内容,这里以 S 表示所有文章,即不考虑文章分级进行数量统计.图 3 中不论级别都基本服从长尾的幂率分布,这一表现与直观认知一致,少量学者占据了多数发文量,这个不平衡数据问题需要充分利用数据构建特征以反映内在模式.

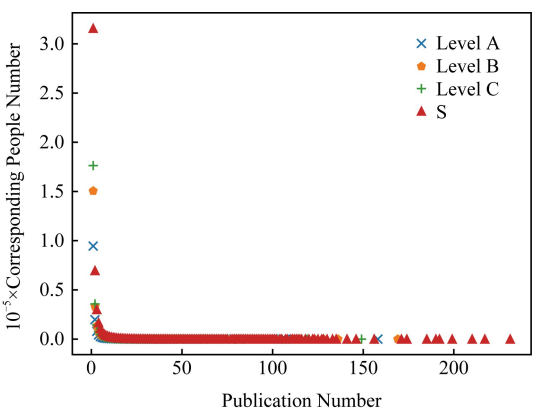


Fig. 3 Publication number and corresponding authors number

图 3 发表篇数与对应人数

将图 3 中发表篇数与对应人数分别取对数得到图 4.

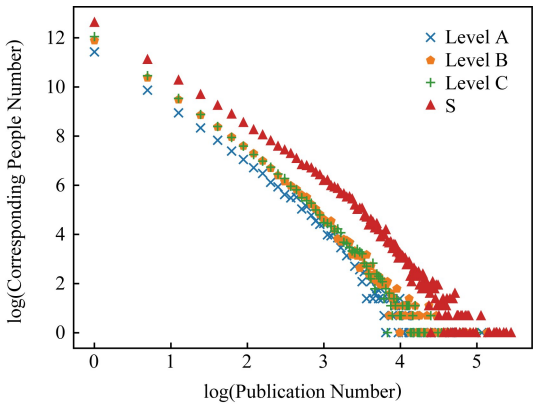


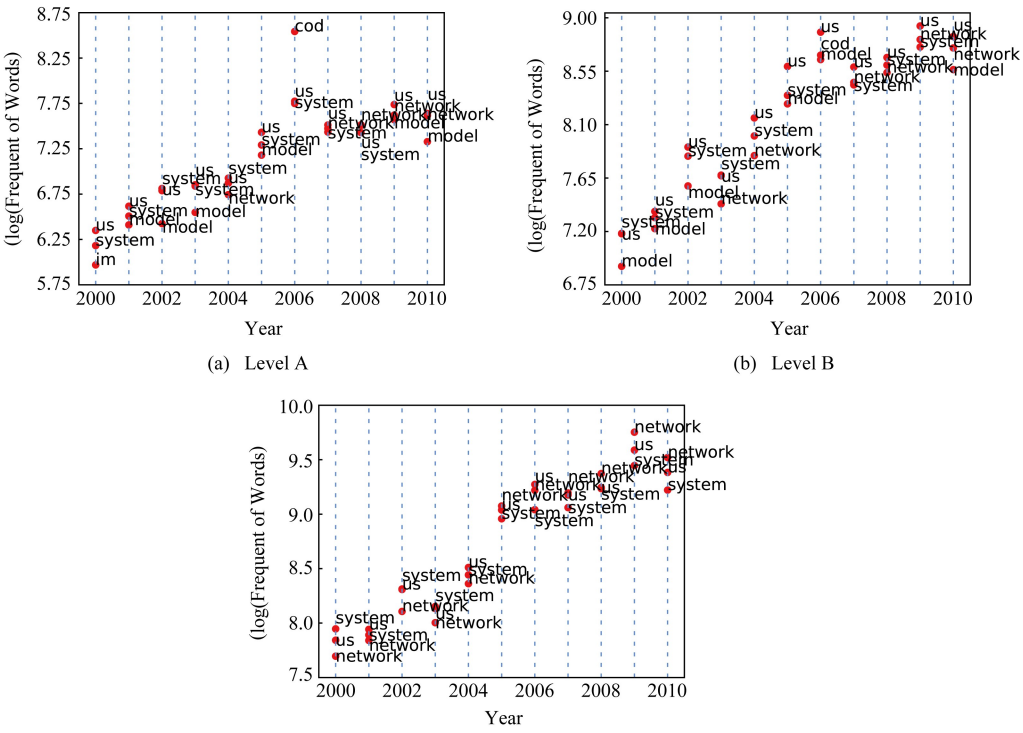
Fig. 4 Logarithms for publication number and corresponding authors number

图 4 发表篇数与对应人数取对数

1) 文章标题.利用文本数据挖掘方法提取各级别所有文章标题中的高频词根.以 2000—2010 年为例,图 5 表示每年前 3 位高频词根随时间变化情况.某些年份词频较为接近,会出现一定程度重叠.图 5 中点的纵坐标与轴对应,表示数值大小,而点右侧词的高低不具有坐标意义,只与相对高低的点对应表示此点的词根.图 5 中不论 A,B,C 类每年高频词根数量持续增长,数量变化趋势基本一致,内容虽大致相同但有少量变化,说明文章标题是一个较为敏感的因素.因此作者文章标题可以作为合作者潜力预测的基本特征.

2) 不同等级中的文章数量.图 6 展示了各等级中,以学者发表文章数量为排序依据,各学者发表不同级别文章数量的分布情况,为了便于表示这里选取排名前 50 位作图.如图 6(a)表示以各学者发表 A 类文章数量为排序依据,前 50 位发表 A,B,C 类文章的分布情况.由图 6 可得,不同排序标准下的排序分布不同,不计级别意义下(图 6(d))的文章数量并不能准确反映作者在各个等级下单独的能力,每个作者在不同等级中都有一定的能力体现,因此文章等级及文章数量可以作为合作者潜力预测的基本特征.

3) 文章发表年份.分别取不同等级中累积发表该类文章数量前 3 位的学者,分析他们发表文章数量的等级分布随时间变化,如图 7 所示.虽然各等级中学者发表文章以该类为主,但仍有其他类的文章发表.学者科研生涯发表文章数量和等级不断变化,考虑年份特征更能反映学者在当下的科研潜力,因此文章的发表年份可以作为合作者潜力预测的基本特征.

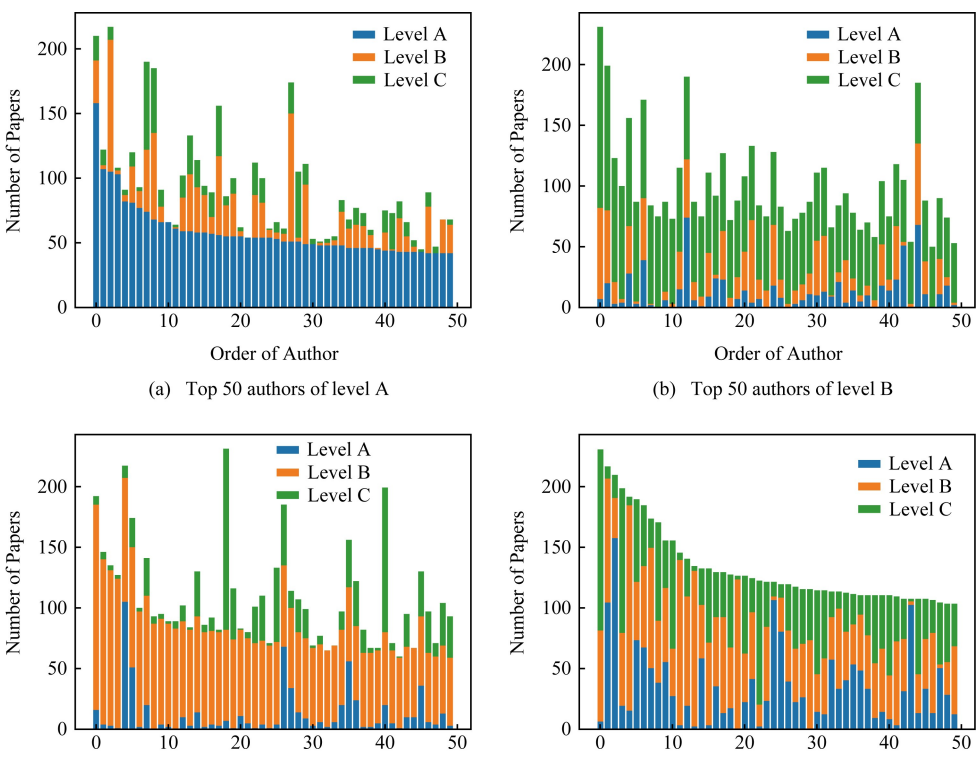


(a) Level A　(b) Level B

(c) Level C

Fig. 5　Top 3 words in titles by year

图 5　文章题目前 3 词汇时间分布

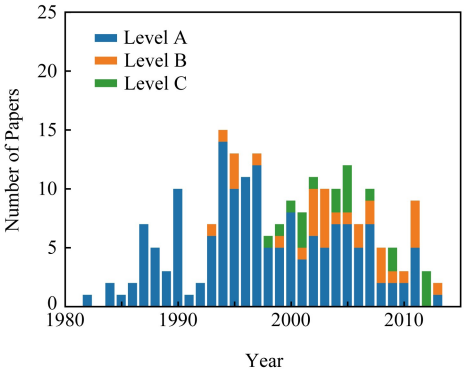


(a) Top 50 authors of level A　(b) Top 50 authors of level B

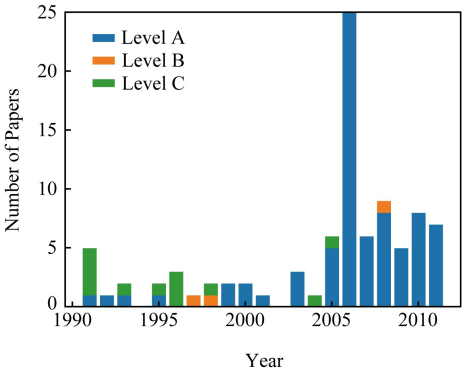
(c) Top 50 authors of level C　(d) Top 50 authors of level S

Fig. 6　Level distribution of top 50 authors in each level

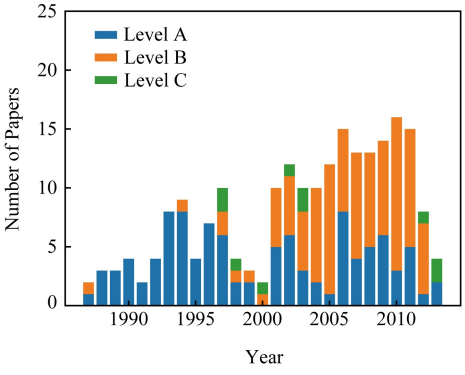
图 6　各等级前 50 位发文等级分布



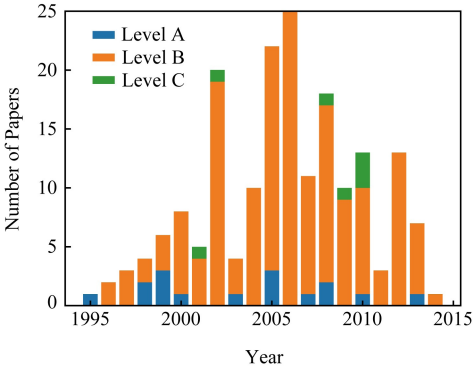
(a) The first author of level A



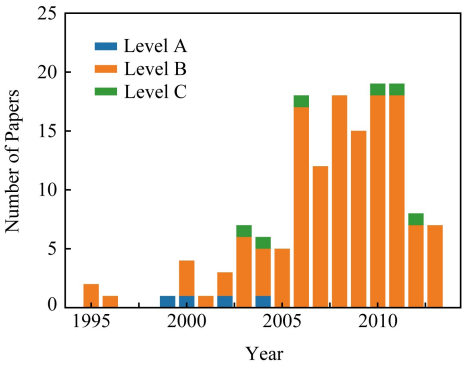
(b) The second author of level A



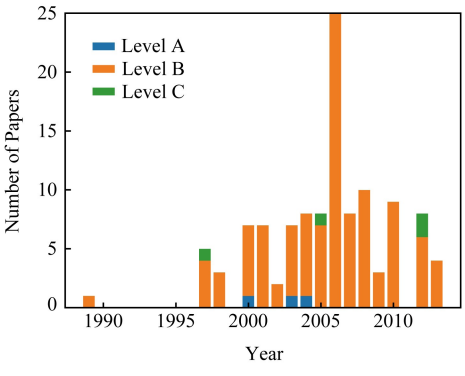
(c) The third author of level A



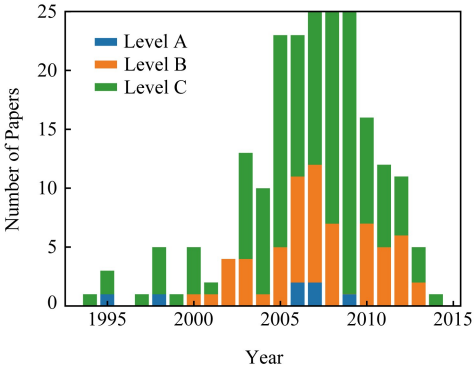
(d) The first author of level B



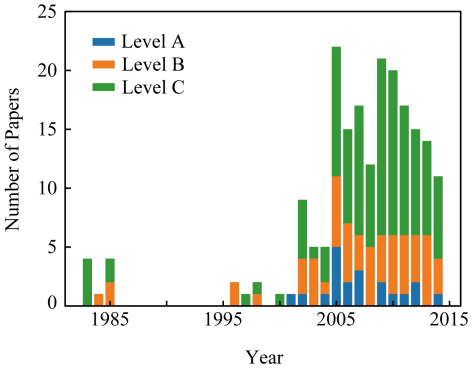
(e) The second author of level B



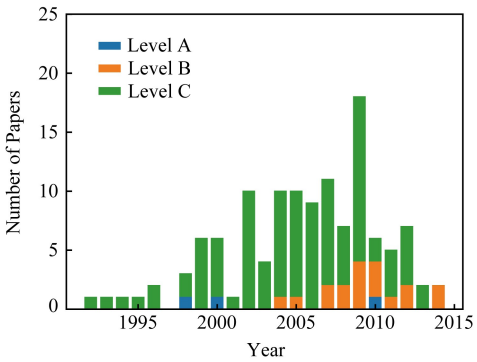
(f) The third author of level B



(g) The first author of level C



(h) The second author of level C



(i)　The third author of level C

Fig. 7　Level distribution by year of top 3 authors in each level

图 7　各等级前 3 位发文等级年份分布

4) 署名序.每篇文章的作者常以合作者列表的形式呈现,每位作者对文章的贡献程度是不一样的,最直观的就是通过学者在作者列表中的位置来反映.例如 1 篇 A 类文章的第 1 作者和第 2 作者比第 2 作者之后的作者对文章的贡献更大,即前 2 位作者较之后的作者有更多 A 类潜力.因此署名序可以作为合作者潜力预测的基本特征.

2.2　样本特征构造

经 2.1 节对学术大数据的分析,提取可用特征:文章标题、文章等级、文章数量、文章发表年份以及署名序.

基于以上基本特征因素进行规范和优化.首先年份特征作为时间因素可以衍生出相关特征:文章发表时间为 t,发表第 1 篇文章距今的时间间隔定义为其学术年龄 AA(academic age)[9];第 1 篇文章的发表时间为 t_0;最近 1 篇文章的发表时间 t_1.对于署名序,由于 1 篇文章的作者列表可能包含多人,进行特征组合会产生大量冗余特征,因此这里只用署名第 2 作者之内的 O_0(包括第 2 作者)的和第 2 作者之外的 O_1 加以区分.而文章级别分为 A,B,C 这 3 类,文章数目计数器记为 N,则优化后的基本特征如图 8 所示:

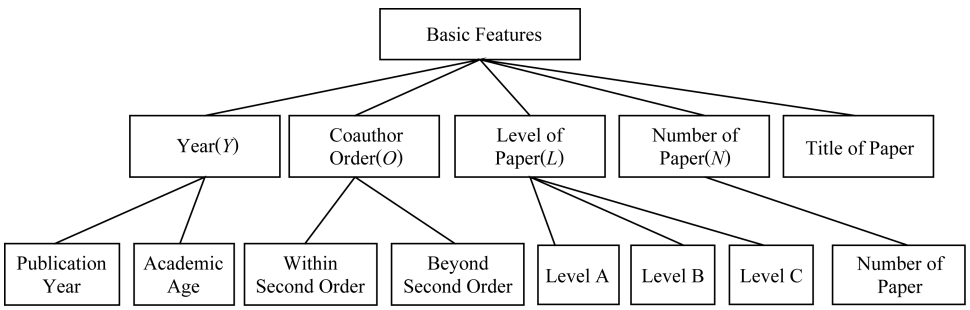


Fig. 8　Basic features after optimizing

图 8　优化后的基本特征

基本特征从简单维度反映了合作者的潜力表现,是预测模型最直接的量化和外在形式.但是仅依据直观意义下的单一统计量不足以适应多样化的合作模式,难以挖掘合作者潜力.因此,对于数字类数据和文本类数据,本文分别采用了不同的特征构造策略:1)以特征工程处理数字类数据,对基本特征进行 1 次组合及 2 次组合,得到较小粒度的可用特征;2)文本类数据特征利用语义信息计算语义相似度构造特征.

1) 数字类数据特征构造.以基本特征为基元扩展特征维度.首先对基本特征进行 1 次组合如表 1 所示,得到特征:NL 特征、OL 特征、YL 特征、NO 特征、NY 特征、YO 特征(所有特征都基于当前合作时间点 t_c 得出).

① NL 特征.不同等级 l_i 中文章的数量 N_{l_i}(l_i 表示文章等级:l_A 表示 A 级别、l_B 表示 B 级别、l_C 表示 C 级别):

$$NL=\{N_{l_i}\mid l_i\in\{l_A,l_B,l_C\}\}. \tag{1}$$

② OL 特征.不同等级 l_i 中文章的署名序 O_{l_i}:

$$OL=\{O_{l_i}\mid l_i\in\{l_A,l_B,l_C\}\}. \tag{2}$$

③ YL 特征.不同时序位置 t_i 和不同等级 l_i 的文章发表时间 $year_{t_i,l_i}$ 或学术年龄 AA_{t_i,l_i}(t_i 表示文章的时序位置:t_0 表示当前学者生涯第 1 篇,t_1 表示当前学者最近 1 篇):

$$YL=\{(year,AA)_{t_i,l_i}\mid t_i\in\{t_0,t_1\},\ l_i\in\{l_A,l_B,l_C\}\}. \tag{3}$$

④ NO 特征.不同署名序 o_i 时的文章数 N_{O_i}(o_i 表示署名序:当作者在文章中署名前 2 位作者时 $o_i=1$,否则 $o_i=0$):

$$NO=\{N_{O_i}\mid o_i\in\{0,1\}\}. \tag{4}$$

⑤ NY 特征.不同时间区间 T_i 中的文章数量 N_{T_i}(T_i 表示文章时间区间的时序位置:$[t_0,t_0+\Delta T)$表示学者学术生涯中第 1 个时间区间、$[t_1-\Delta T,t_1)$表示最近 1 个时间区间,其中 ΔT 表示时间区间长度,如 $\Delta T=5$,则统计学术生涯前 5 年和最近 5 年中的文章数量):

$$NY=\{N_{T_i}\mid T_i\in\{[t_0,t_0+\Delta T),\ [t_1-\Delta T,t_1)\}\}. \tag{5}$$

⑥ YO 特征.不同时序位置 t_i 和不同署名序 o_i 的文章发表时间 $year_{t_i,o_i}$ 或学术年龄 AA_{t_i,o_i}:

$$YO=\{(year,AA)_{t_i,o_i}\mid t_i\in\{t_0,t_1\},\ o_i\in\{0,1\}\}. \tag{6}$$

Table 1 Combining Basic Features of Papers

表 1 文章基本特征组合

Basic Features	Basic Features			
	L	N	O	Y
L		NL	OL	YL
N	NL		NO	NY
O	OL	NO		YO
Y	YL	NY	YO	

细化特征粒度,排除意义重复特征,对以上特征二次组合得到 $YL\&O$ 特征和 $NO\&YL$ 特征.

⑦ $YL\&O$ 特征.不同时序位置 t_i、不同署名序 o_i、不同等级 l_i 的文章发表年份 $year_{t_i,o_i,l_i}$ 或学术年龄 AA_{t_i,o_i,l_i}:

$$YL\&O=\{(year,AA)_{t_i,o_i,l_i}\mid t_i\in\{t_0,t_1\},\ o_i\in\{0,1\},l_i\in\{l_A,l_B,l_C\}\}. \tag{7}$$

⑧ $NO\&YL$ 特征.不同时间区间 T_i、不同署名序 o_i、不同等级 l_i 的文章数量 N_{T_i,o_i,l_i}:

$$NO\&YL=\{N_{T_i,o_i,l_i}\mid T_i\in\{[t_0,t_0+\Delta T),\ [t_1-\Delta T)\},o_i\in\{0,1\},l_i\in\{l_A,l_B,l_C\}\}. \tag{8}$$

2) 文本类数据特征构造.文本数据不同于数字类数据.每个文本在形式上由包括标点在内的字符组成,由词到句,由句到篇.不论是在文本的自底向上或自顶向下的层次解析中,形式相同的一段字符串在不同的语境下可得到不同的含义.文本的一致性和多义性决定了其独特的处理方式.因此本文利用潜在语义索引[20]方法,计算基于集成语料库子空间的标题特征值以及标题间相似度.

首先给出 3 个定义:

定义 1. 文章标题全集 $sumTitle(T^*,L^*)$.

$$sumTitle(T^*,L^*)=\bigcup\{title_{p_i}\mid year_{p_i}\in T^*,l_{p_i}=L^*\}, \tag{9}$$

其中,T^* 表示发表年份区间,L^* 表示文章等级(标题 $title_{p_i}$、年 $year_{p_i}$、级别 l_{p_i} 对应于同一篇文章 p_i).

定义 2. 文本特征值计算函数 $Cal(x,X)$.其中,x 为待计算的文本样本,X 为对应的计算子空间.

定义 3. 文本相似度运算函数 $dis(y_i,y_j)$.其中 y_i,y_j 分别为文本合集.

基于文本定义可得 2 个文本特征:

1) TIT 特征.分别以作者 au 第 1 篇 t_0、最近 1 篇 t_1 及到当前时间 t_c 为止累积发表文章标题合集作为计算样本 x,以样本 x 中文章在当时年份 t_p、当时年份之前 1 个时间区间 $[t_p-\Delta T,t_p)$及到当时年份为止 $[0,t_p)$时间段的文章标题全集分别为计算子空间 X(计算子空间 X 中所有文章等级与计算样本 x 的等级 l_p 一致),将样本和子空间代入特征值函数得出标题特征:

$$\begin{cases}x=sumTitle_{au}(t_p=\{t_0,t_1,[0,t_c)\},\ l_p=\{l_A,l_B,l_C\}),\\ X=sumTitle_{\text{domain}}(\{t_p,[t_p-\Delta T,t_p),\ [0,t_p)\},l_p),\\ TIT=\{Cal(x_i,X_i)\mid x_i\in x,X_i\in X\}.\end{cases} \tag{10}$$

2) $SIM_{\text{title}}(au_i,au_j)$相似度.合作者 au_i 和 au_j 到当时年份为止 $[0,t_p)$在各等级上文章标题全集之间的文本相似度:

$$\begin{cases}y_i=sumTitle_{au_i}(t_p=[0,t_c),l_{p_1}=\{l_A,l_B,l_C\}),\\ y_j=sumTitle_{au_j}(t_p=[0,t_c),l_{p_2}=l_{p_1}),\\ SIM_{\text{title}}(au_i,au_j)=dis(y_i,y_j).\end{cases} \tag{11}$$

2.3 集成学习方法

集成分类器如图 9 所示,利用多个基学习器参与学习,通过投票或平均选择最适应当前任务的分类器,提高泛化性能.

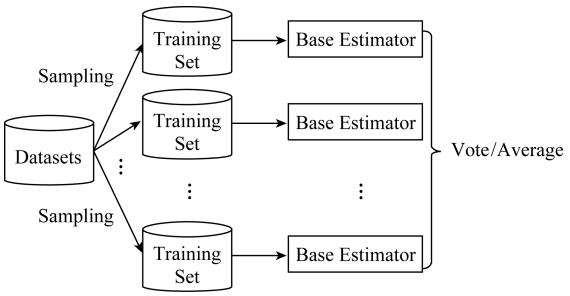


Fig. 9　Ensemble estimators

图 9　集成分类器

所用集成学习方法简述:

1) AdaBoost[14].该方法先从初始训练集训练出1个基学习器,再根据基学习器的表现对训练样本进行调整,使得先前基学习器的错分样本能够在后续中得到更多的训练,基于调整后的样本分布进行下一个基学习器的训练.如此重复直到基学习器数量达到预先指定的值,最终将这些基学习器进行加权结合.

2) Bagging[15].该方法首先经过多次随机有放回采样,得到多个采样集.使得有的样本多次出现,有的样本则从未出现.进而个体学习器之间既有差异又能进行有效学习.之后从每个采样集中训练出1个基学习器,最终将这些基学习器进行结合.

3) 随机森林[16](random forest, RF).在以决策树为基学习器的 Bagging 集成的基础上,在训练过程中引入随机属性选择,不同于传统的决策树在当前的属性全集中选择1个最优属性,而是从属性集合中随机选择1个属性的子集,再从子集中选1个最优属性用于划分.

3　实验设计

3.1　数据描述与实验设置

ArnetMiner 数据集上有丰富的文章和作者信息,包含文章数据 2 092 356 条、作者数据 1 712 433 条、文章作者匹配数据 5 192 998 条.每条文章数据包括 ID、题目、作者、年份、所发机构、期刊等;每条作者数据包括 ID、姓名、机构、研究兴趣等;文章作者匹配数据通过各自的 ID 把文章和作者联系起来.ArnetMiner 数据的优势在于给每个作者赋予了唯一的 ID,使重名消歧的问题从数据源头得以解决.《CCF 推荐目录》中包含计算机 10 个领域的 600 多个各类期刊会议.本文将 ArnetMiner 中《CCF 推荐目录》的全部数据抽出,根据《CCF 推荐目录》的论文分级从 ArnetMiner 中抽取计算机领域整个数据集作为研究对象,以期刊和会议的等级高低作为文章的级别标签(其中 A 类文章 100 324 篇,B 类文章 162 634 篇,C 类文章 208 422 篇,总计 471 380 篇),得到计算机领域的全部数据进行实验.

从以上真实数据中抽取合作边,根据第 2 节所述合作者潜力预测模型中的式(1)～(11)构建样本特征,结果等级为 A 类文章样本标签记为"1",否则样本标签记为"0".依流程图 1 把每个合作边样本构造为〈特征,标签〉样本,得到样本数据集.为了兼顾模型的准确率和运行的时间开销,本文的实验以当前年份之前 10 年数据为训练集,当前年份之后 3 年数据为测试集进行验证.如当前年份为 2010 年时,训练集为 2000—2010 年的样本,测试集为 2001—2003 年的样本.

3.2　评价指标

本文选用机器学习分类常用的准确率、召回率、F1 分数和模型学习时间作为评价模型的指标.假定某类标签的预测集合为 PS.根据预测标签和真实标签可以将 PS 分为表 2 混淆矩阵中的 4 组.

Table 2　Confused Matrix

表 2　混淆矩阵

Labels	Labels	
	Actual is True	Actual is False
Predicted is True	True Positive(TP)	False Positive(FP)
Predicted is False	False Negative(FN)	True Negative(TN)

TP:真实标签为正例被正确判定为正例;

FN:真实标签为正例未被正确判定为正例;

FP:真实标签为负例的被错误判定为正例;

TN:真实标签为负例的未被判定为正例.

由混淆矩阵得到准确率(precision, P)、召回率(recall, R)、F1 分数(F1-score, F)计算为

$$P=\frac{TP}{TP+FP}, \tag{12}$$

$$R=\frac{TP}{TP+FN}, \tag{13}$$

$$F=\frac{2\times P\times R}{P+R}. \tag{14}$$

3.3　实验结果及分析

为了验证本文构造模型在合作者潜力预测问题中的适应性,设置实验对 TP,FN,FP,TN 四个指标进行测试比较.同时引入决策树(decision tree,

DT)、K-近邻(K-nearest neighbor, KNN)、逻辑回归(logistic regression, LR)和支持向量机(support vector machine, SVM)多种传统学习算法作为基于集成学习的科研合作者潜力预测模型的对比方法,进一步验证其有效性.

图 10 以 10 年全部数据中不同百分比的样本作为训练集,集成学习方法基分类器个数为 300 时,Adaboost,Bagging,RF,DT,KNN,LR,SVM 多种类型的算法在 4 个指标的实验性能(其中 SVM 时间开销巨大,因此图 10(d)只展示了除 SVM 外的 6 种算法运行时间,同时 DT 和 LR 运行较快,曲线基本重叠).结果显示,集成学习算法虽然时间开销较高,但是准确率、召回率和 F1 分数都远高于对照算法.同时,本文基于集成学习算法的模型在较小的训练集时就已经能取得较好的效果,即以较少的数据量快速收敛于较高的性能.其中 Bagging 对模型的适应性最好,但是运行时间更长.运行时间、性能参数与数据量正相关,但是使用 20%的训练集样本就基本接近性能最优值,此时的运算时间较低,因此时间开销并不大.

图 11 为以 10 年中全部数据作为训练集,增加集成学习中基分类器个数时 Adaboost,Bagging,

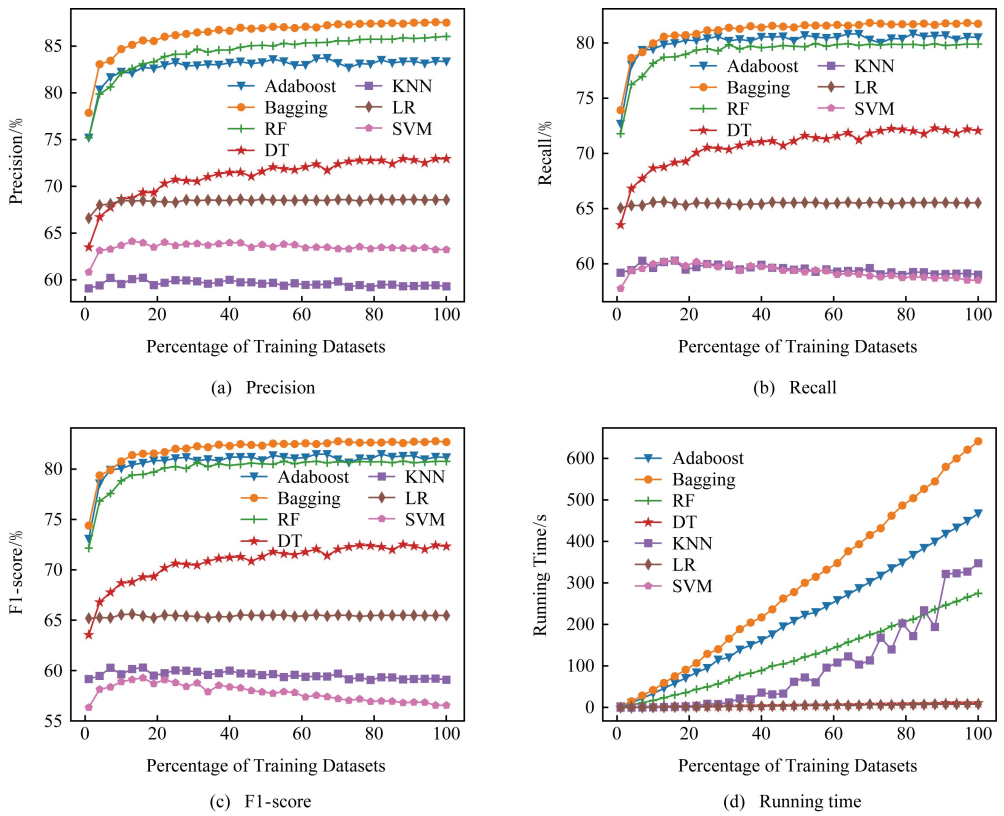
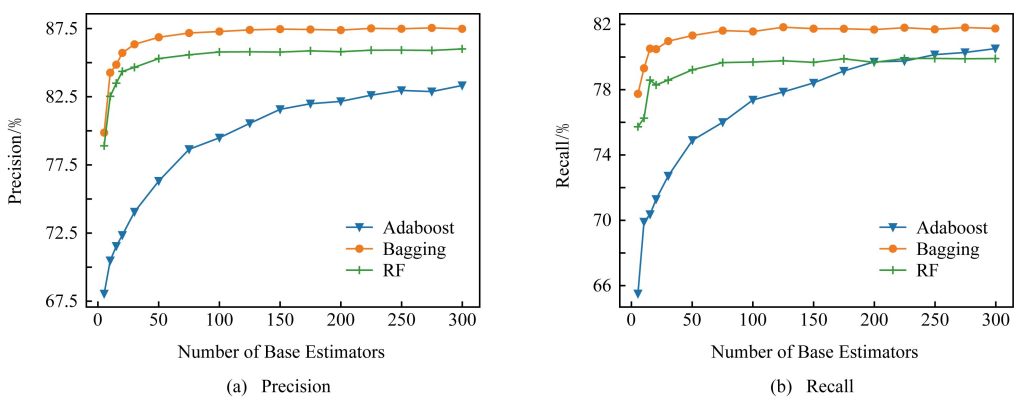
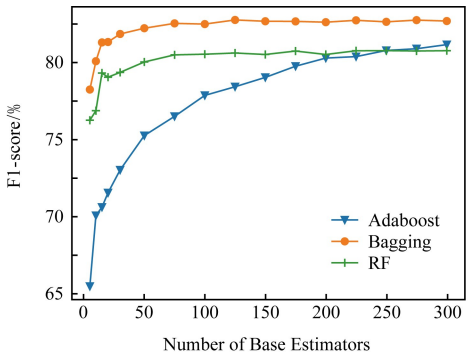


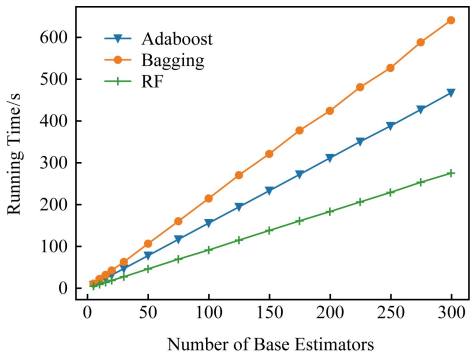
Fig. 10 Experimental results with training dataset increasing

图 10 训练集增加的实验效果





(c) F1-score



(d) Running time

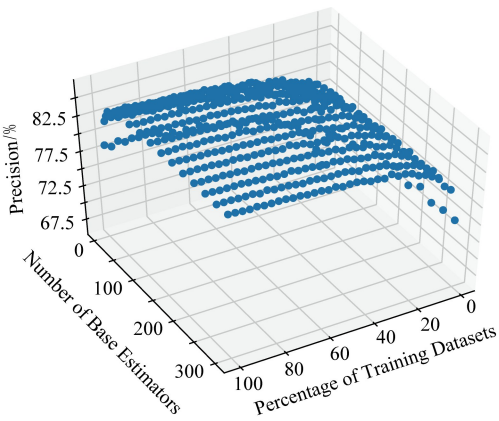
Fig. 11　Experimental results with estimator number increasing

图 11　基分类器增加的实验效果

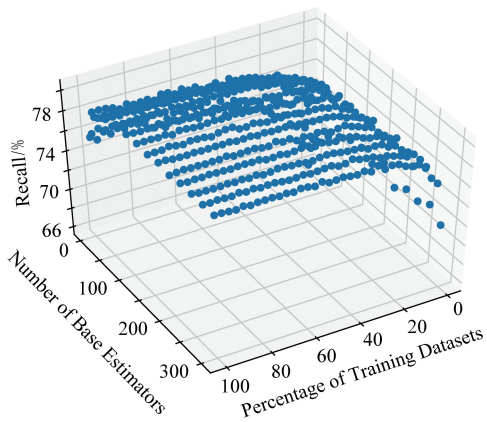
RF 这 3 种算法在 4 个指标的实验性能.与增加训练集数据时类似,本文模型在较少的基训练器时基本接近大量基训练器的效果.而运行时间主要取决于集成学习方法自身的复杂度,基本呈线性分布.

3 种集成学习算法的性能表现趋势大致相同,因此图 12 以 RF 为代表,用三维散点图的形式表示训练集和基分类器同时增加时的实验效果.更直观地说明了增加训练集和基分类器个数对实验性能的影响.

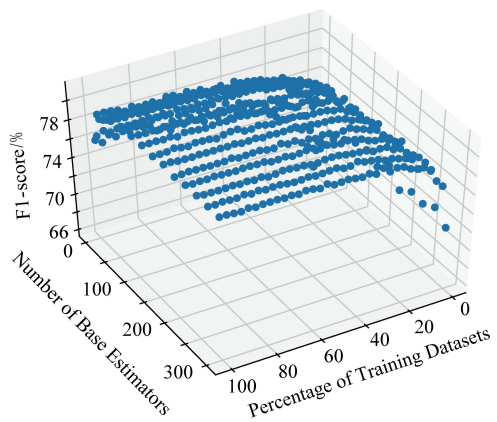
综合上述实验结果可得:1)3 种经典集成学习方法的准确率、召回率和 F1 分数都超过了 0.8,较好地完成了合作者的潜力预测问题.Bagging 算法最



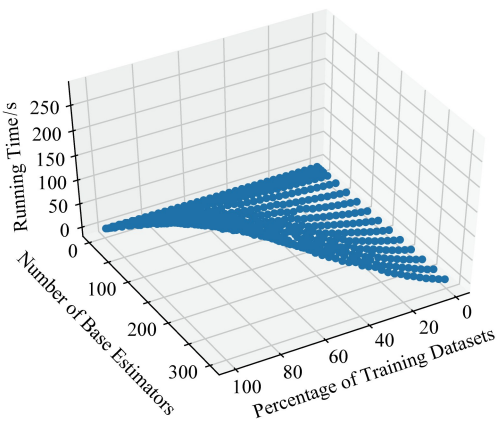
(a) Precision



(b) Recall



(c) F1-score



(d) Running time

Fig.12　Experimental results with training dataset and estimators number increasing

图 12　训练集和基分类器增加时 RF 的实验效果

能适应本文所提模型,准确率、召回率、F1 分数分别达到 87%,82%,82%.RF 虽然性能略差,但是运行时间最快;2)模型的样本收敛较快,少量训练集时性能参数基本达到最优;3)基分类器数量较少时性能参数基本达到最优;4)以上 2 点保证了模型较低的时间开销.

4 结束语

基于文章等级与合作者属性相关这一假设,本文研究了大数据背景下的合作者潜力预测问题,从大量合作关系中挖掘不同等级的合作表现,指导学者进行合作者选择.为了训练和评估模型,本文从学术大数据中抽取并构造了一系列特征来描述合作者潜力,把 ArnetMiner 的文章和学者信息与《CCF 推荐目录》匹配构造包含等级的文章数据集和包含等级的学者数据集作为样本集.同时定义了一系列学术背景下的学者个人特征描述及学者间相关性特征描述,并将经典集成学习方法应用于所构造的样本.实验结果说明了本文所提模型的实用性和优越性,从而可以为学者选择有潜力合作者提供参考性意见,有助于个人科研效率最大化.

本文未来的工作将继续拓展特征的丰富性来更全面地刻画合作者潜力,以《CCF 推荐目录》中的不同分级为标准,向多维数据扩展,如期刊、会议、作者主页,爬取完整数据,提升模型性能,进一步挖掘合作模式.

参 考 文 献

[1] Fortunato S, Bergstrom C T, Börner K, et al. Science of science [J]. Science, 2018, 359(6379): 185-185

[2] Lee S, Bozeman B. The impact of research collaboration on scientific productivity [J]. Social Studies of Science, 2005, 35(5): 673-702

[3] Newman M E. The structure of scientific collaboration networks [J]. Proceedings of the National Academy of Sciences of the United States of America, 2001, 98(2): 404-409

[4] Michele A, Moro M, Lopes G R. Using link semantics to recommend collaborations in academic social networks [C] //Proc of the 5th Workshop on Simplifying Complex Networks for Practitioners-Simplex. New York: ACM, 2013: 833-840

[5] Xia Feng, Chen Zhen, Wang Wei, et al. MVCWalker: Random walk-based most valuable collaborators recommendation exploiting academic factors [J]. IEEE Transactions on Emerging Topics in Computing, 2014, 2(3): 364-375

[6] Tang Jie, Wu Sen, Sun Jimeng, et al. Cross-domain collaboration recommendation [C] //Proc of the 18th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2012: 1285-1293

[7] Su Xiaoyan, Wang Wei, Yu Shuo, et al. Can academic conferences promote research collaboration [C] //Proc of the 16th ACM/IEEE-CS on Joint Conf on Digital Libraries. New York: ACM, 2016: 231-232

[8] Lopes G R, Moro M M, Wives L K, et al. Collaboration recommendation on academic social networks [C] //Proc of the 29th Int Conf on Conceptual Modeling. Berlin: Springer, 2010: 190-199

[9] Wang Wei, Liu Jiaying, Xia Feng, et al. Shifu: Deep learning based advisor-advisee relationship mining in scholarly big data [C] //Proc of the 26th Int Conf on World Wide Web Companion. Lyon, France: International World Wide Web Conferences Committee (IW3C2), 2017: 303-310

[10] Sagiroglu S, Sinanc D. Big data: A review [C] //Proc of the 9th Int Conf on Collaboration Technologies and Systems. Piscataway, NJ: IEEE, 2013: 42-47

[11] Xia Feng, Wang Wei, Bekele T M, et al. Big scholarly data: A survey [J]. IEEE Transactions on Big Data, 2017, 3(1): 18-35

[12] Williams K, Wu Jian, Choudhury S R, et al. Scholarly big data information extraction and integration in the CiteSeerX digital library [C] //Proc of the 33rd Int Conf on Data Engineering Workshops. Piscataway, NJ: IEEE, 2017: 68-73

[13] Ditterrich T G. Machine learning research: Four current direction [J]. Artificial Intelligence Magazine, 1997, 18(4): 97-136

[14] Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting [J]. Journal of Computer and System Sciences, 1997, 55(1): 119-139

[15] Breiman L. Bagging predictors [J]. Machine Learning, 1996, 24(2): 123-140

[16] Breiman L. Random forests [J]. Machine Learning, 2001, 45(1): 5-32

[17] Qian Suchi, Peng Furong, Li Xiang, et al. The hierarchical model to Ali mobile recommendation competition [C] //Proc of the 15th Int Conf on Data Mining Workshop. Piscataway, NJ: IEEE, 2015: 1070-1077

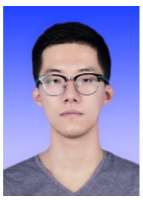
[18] Hu Kaixian, Liang Ying, Xu Hongbo, et al. Amethod for social network user identify feature recognition [J]. Journal of Computer Research and Development, 2016, 53(11): 2630-2644 (in Chinese)

(胡开先,梁英,许洪波,等. 一种社会网络用户身份特征识别方法[J]. 计算机研究与发展,2016,53(11):2630-2644)

[19] Tang Jie, Zhang Jing, Yao Limin, et al. ArnetMiner: Extraction and mining of academic social networks [C] //Proc of the 12th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2008: 990-998

[20] Deerwester S. Indexing by latent semantic analysis [J]. Journal of the Association for Information Science & Technology, 2010, 41(6): 391-407

[21] Sonnenwald D H. Scientific collaboration [J]. Annual Review of Information Science and Technology, 2007, 41(1): 643-681



Ai Ke, born in 1992. Master. His main research interests include data mining and complex networks.

Ma Guoshuai, born in 1992. PhD. His main research interests include data mining and link prediction.

Yang Kaikai, born in 1993. Master. Her main research interests include data mining and complex networks.

Qian Yuhua, born in 1976. Professor and PhD supervisor. Member of CCF. His main research interests include granular computing, social computing and machine learning.