基于 Stacking 模型的学术论文多标签分类系统构建

刘爱琴　郭少鹏

摘　要　学术论文高质量多标签自动分类是推动学术研究发展的关键程序之一。本研究利用 Stacking 模型将随机森林、支持向量机、极限树、极端梯度提升和神经网络五个分类器融合为一个异质集成分类器，并利用基于问题转换思想的多二分类模型将该分类器应用于学术论文多标签分类。根据学术论文的特点，依次实现了与之配套的论文特征提取模块、TF-IDF 加权模块、数据预处理模块，最终构建成一个面向学术论文的多标签分类系统。仿真实验验证了本研究构建的学术论文多标签分类系统在处理学术论文多标签分类问题时，较传统的单模型分类器或同质集成模型分类器在泛化能力、稳定性与准确率方面都有一定程度的提升。图 9。参考文献 21。

关键词　论文分类　Stacking 模型　多标签分类　多二分类模型

Construction of Multi-Label Classification System for Academic Papers Based on Stacking Model

Liu Aiqin　Guo Shaopeng

Abstract: High-quality multi-label automatic classification of academic papers is a key step to promote the development of academic research. In this study, the five classifiers of random forest, support vector machine, limit tree, extreme gradient boosting, and neural network are fused into a heterogeneous ensemble classifier using Stacking model, and the multi-binary classification model based on problem transformation idea is used to apply the classifier to multi-label classification of academic papers. According to the characteristics of academic papers, the supporting paper feature extraction module, TF-IDF weighting module and data pre-processing module are realized in turn, and finally a multi-label classification system for academic papers is constructed. Simulation experiment verifies that the multi-label classification system for academic papers constructed in this study has a certain degree of improvement in generalization ability, stability and accuracy compared with the traditional single-model classifier or homogeneous ensemble model classifier for the multi-label classification problem of academic papers. 9 figs. 21 refs.

Keywords: Paper Classification; Stacking Model; Multi-Label Classification; Multi-Binary Classification Model

1　研究背景

学术论文是科研人员进行研究发现和发表学术观点的主要形式[1]，具有针对性强、文本内容庞杂且专业词汇更新速度快等特点。快速从学术论文中提取出与研究相关且有价值的信息，是降低科研人员知识发现成本的关键。在当前文本分析环境下，传统的人工检阅模式成本较高且具有一定的主观性[2]。在海量的数字文献资源中，学术论文分类是实现精准文献检索、推荐和文献计量分析的基础[3]。

算力的提升为文本分类的研究发展提供了不可或缺的资源体系支撑，国内外学者也在文本分类领域取得了一定的研究成果，提出了一些如分类器链算法和二进制相关算法等多标签分类算法。Amanda 和 Ross[4] 提出了基于决策树算法改进的 ML-DT 算法。Elisseeff 和 Weston[5] 基于支持向量机(SVM)算法提出了一种改进的 Rank-SVM 算法。Schapire 和 Singer[6] 提出了一种 Boosting 算法，能够通过训练学习有效提高样本的分类精度。随着深度学习技术的发展，Zhang 和 Zhou[7] 使用全连接神

DOI:10.13666/j.cnki.jnlc.2024.0210

经网络来解决多标签文本分类问题,并提出了基于KNN算法改进的ML-KNN算法[8,9]。李楚贞等[10]结合卷积神经网络,提出一种新的复杂文本多标签分类算法,能够获取较高精度的分类结果。刘爱琴和马小宁[11]通过研究概率主题模型,提出了一种有助于知识快速聚类的短文本自动分类系统。邵孟良和齐德昱[12]提出一种非独立同分布的多实例多标签分类算法,可以很好地解决Boosting算法计算成本高、学习时间长的问题。马雨萌等[13]提出了一种融合BERT模型和多尺度CNN模型的多标签分类方法,实现了科技政策文本内容的自动编码与多主题分类。

上述研究方法为学术论文多标签分类提供了崭新的研究思路,但其预测结果很大程度上依赖于具象化问题,其场景适应性、泛化能力、模型稳定性均有待提升。而Stacking集成学习算法能够以一个元分类器模型为基础进行整合重组,形成一个强分类器,从而提升分类系统的泛化能力。史佳琪和张建华[14]提出了一种基于多模型融合Stacking集成学习方式的负荷预测方法,与传统单模型预测相比,该方法有着较高的预测精度。李寿山和黄居仁[15]利用Stacking模型进行组合分类,发现该组合方法在所有领域都能够获得比参考基分类方法更好的分类效果,从而克服了分类方法领域的依赖困境。李珩等[16]构造了叠加式框架结构,组合4种分类器,并融合各种可能的上下文信息作为各层分类器的输入特征向量,发现组合后的分类器无论在准确率还是召回率上都有所提高。王彦莹等[17]利用Stacking集成学习思想,构建了Stacking-TRN-SG模型,大大提高了历史古籍识别模型的识别准确度,提高了根据历史古籍构建知识图谱的效率。

综上所述,现有对于多标签分类模型的研究大部分为单算法,仅有少数使用了同质集成算法,并且鲜有针对学术论文多标签分类的研究。学术论文所覆盖的学科范围较广,拥有庞大的数量集,但现有模型均存在着不同程度的过拟合或欠拟合现象,分类准确率、稳定性、普适性等方面的表现难以满足学术论文多标签分类对算法性能的要求。鉴于此,本研究在学术论文语义关联的基础上,构建了基于Stacking模型的学术论文多标签分类系统,旨在克服传统分类器存在的泛化能力弱、稳定性低等问题,进而提高学术论文多标签分类的精准度,提高实际工作中的检索效率,并为相关学者进行学术论文多标签分类研究提供参考。

2 基于Stacking模型的学术论文多标签分类系统构建

2.1 Stacking模型理论基础

Stacking模型的中文名为“堆叠”模型,是一种将多个异质分类器融合为一个集成分类器的模型,在机器学习应用中有着出色的表现。经典的Stacking模型由基层和元层两层模型堆叠而成。基层由多个基分类器组成,各基分类器分别对数据进行学习并生成预测结果;元层通常只有一个元分类器,以各基分类器输出的预测结果作为特征项进行学习,并输出整个模型的分类预测结果。由Stacking模型融合而成的异质集成分类器能有效克服单算法存在的过拟合问题,提高分类的准确率、稳定性和普适性,其综合分类性能相较于子分类器有着较大程度的提升。

Stacking模型在其基分类器学习过程中巧妙地嵌入了K折交叉验证方法。首先,对训练集进行分割,设置分割份数为K;其次,将K-1份作为训练集进行学习;最后,对剩余的1份进行预测。这样循环K次后即完成了对所有训练集数据的预测。此方法可有效降低模型的过拟合风险,提高分类器的稳健性与可靠性。

Stacking模型的分类性能取决于各基分类器与元分类器的选取,选取的原则是“好而不同”,即各子分类器在具备较好性能的同时又要保证内部差异性。本研究对十几种分类算法的分类性能进行反复测试,兼顾效率因素,综合研判后选取了随机森林、支持向量机、极限树、极端梯度提升四种分类算法作为基分类器,神经网络作为元分类器,模型内部交叉验证学习的折数设置为10折。

2.2 系统构建

基于 Stacking 模型的学术论文多标签分类系统由“训练系统”和“预测系统”两大子系统构成，两大子系统又由“论文特征提取模块”、“TF-IDF 加权模块”、“数据预处理模块”和“Stacking 模型分类模块”有机组成。

2.2.1 论文特征提取模块

(1)选取视图

现有研究构建的学术论文自动分类系统往往从单一视图出发(如:标题、关键词或摘要)，但由于学术论文的文本长度及专业词汇数量等不断增加，从多个视图方向实现学术论文自动分类是当下文本自动分类研究的关键所在。跨学科合作在应对复杂问题中发挥着愈加重要的作用[18]，若仅以单视图选取论文特征词则难以充分挖掘论文所属类别。因此，本研究设计了以 1∶1∶1 的比例从标题、关键词和摘要三大主视图中提取论文特征词的方案。

(2)去停用词

停用词指的是在单篇论文中出现频率较高且在多领域论文中都有涉及，但自身并不代表任何实义的词汇。停用词对论文表达或理解并无作用，对论文主题更无所裨益。因此，需要引用停用词表删去无实意词汇[19]。本研究将《百度停用词表》《四川大学机器智能实验室停用词库》《中文停用词表》《哈工大停用词表》四大主流停用词表取并集作为停用词。

(3)中文分词

中文与英文不同，不存在天然的分隔符来限定词的范围[20]。因此，从中文汉字串中切分出表达真正意义的词是论文特征提取模块最关键的步骤，当前主流的分词原理仍然是基于词典的方法。本研究调用 python 语言中的 jieba 分词库，其在分词的准确性与速度方面均有着良好的性能。jieba 分词库中的精准分词模式会尽可能精确地将句子切分成表达实义的词，特别适合用于文本分类所需的特征词提取。

(4)文本表示模型

文本表示模型是指将学术论文中的文字文本进行形式转换，将非结构化的论文文本转化为结构化模式，实现计算机系统的自动辨别分类。词是在文本分类过程中使用较为广泛的特征项。向量空间模型(Vector Space Model)因对特征覆盖性较全而得到较为广泛的使用[21]。向量空间模型将论文文本中的每个特征词分别作为一个维度，并以数字形式代表特征词在该论文中的重要程度，如此即可将非结构化的论文表示为一条由 n 个特征词构成的 n 维空间中的向量。

(5)标签处理

本研究所使用的多标签分类模型为多二分类模型，与文本表示的向量空间模型类似，需要将论文的每个类别分别视作一个维度，由所有的类别组成一个多维空间，并统计每篇论文在各个类别下的值(属于该类别值为 1，不属于则为 0)，生成各类别的空间向量。学术论文的多标签分类问题由此转化为依次判断该论文是否属于每个类别的多二分类问题。

2.2.2 TF-IDF 加权模块

构建向量空间模型后，如何用数值表示每个特征词对某篇论文的重要程度是最关键的问题。本研究使用了当前领域内主流的加权方式 TF-IDF(Term Frequency-Inverse Document Frequency)框架作为加权方法。TF-IDF 被广泛应用于数据挖掘与信息检索之中，其中 TF 指词频，IDF 代表逆文本频率指数。TF-IDF 框架作为加权方法是指当特征词在某篇学术论文中的出现频次较高(TF)，且该特征词在其他论文中出现的频率较低(DF)时，该特征词被视为于该篇学术论文而言具有较强的区分特征，可以用于该篇论文的分类。TF-IDF 可用算式表示为 TF * IDF，TF 指特征词 t 在文档 d 中出现的频率，IDF 表示特征词 t 在所有论文中出现次数的倒数，TF * IDF 的结果可以用于表示特征词 t 对于文档 d 的区分程度，即结果越大，区分能力越强。具体加权流程如下：

(1)依据论文特征提取模块中生成的特征词列表，统计每个特征词在论文训练集中出现的频数，依据公式(1)计算每个特征词的 IDF 值(其中 N 为训练集论文的数量，n 为出现过该词的论文

数量,L 为常数,取 0.01),并生成一个《特征词 IDF 表》供待分类集加权使用。

$$IDF(T_K)=\log\left(\frac{N}{n_k}+L\right) \qquad 公式(1)$$

(2)统计训练集中每篇论文的每个特征词在该篇论文中的题目、关键词和摘要中出现的频数,依据公式(2)计算每个特征词的 TF 值,并依据公式(3)将该特征词的 TF 值与对应的 IDF 值相乘得出权重指数。

$$TF_d(i)=\frac{Fre_d(i)}{\sum_{j=1}^{n}Fre_d(j)} \qquad 公式(2)$$

$$Weight(T_{ij})=TF(jj)\times IDF(j) \qquad 公式(3)$$

(3)在预测子系统中,结合步骤(1)生成的《特征词 IDF 表》重复步骤(2),对分类集的论文特征词进行加权。

2.2.3 数据预处理模块

对数据集进行预处理可以提高论文分类质量并缩短分类时间。根据学术论文多标签分类的特点,数据预处理的主要部分是数据集成与数据规约。

(1)数据集成。训练子系统的前两个模块中生成的分别是训练集论文的特征词空间向量数据集和类别空间向量数据集。但两个数据集不能直接用于分类器训练,需将其整合为一个多标签训练数据集。

(2)数据规约。为提高分类系统运行效率,在尽可能降低对分类效果影响程度的前提下,需要将数据集进行压缩表示。文本分类数据集的共同特点是维数巨大,严重影响分类器效率。同时,虽然在特征提取阶段进行了停用词过滤,但仍会存在部分区分能力较弱的词,因此需要对学术论文分类数据进行维度规约。

维度规约的主流方法有基础的方差分析、相关性分析等,也有进阶的小波变换、主成分分析等。本研究对以上四种维度规约方法进行反复实验分析处理,结果显示:对于区分能力较弱的特征词,方差分析的发现作用不明显且分类效果受到了严重影响;文本分类数据的特征项并非固定,而是由大量的训练集确定的,小波变换和主成分分析会使得原本的实义特征项转变为没有实际意义的特征项,因此难以适用于学术论文分类的数据集降维;相关性分析在大幅减少数据维度的基础上,并未严重影响分类效果,降维效果较为理想。

维度规约的核心思想是判断一个特征项与类别的相关度,从本质上发现弱区分能力的特征词。基于多标签分类的特点,本研究改进了传统的相关性分析方法。首先,将特征词同所有类别依次进行相关性分析;其次,取最大值作为该特征词的相关度;最后,通过设置阈值来排除弱区分能力的特征词。多次实验表明,设置阈值为 0.3 时效果最佳,在降维后会生成最终的特征词列表,保存该列表作为预测子系统中表示待分类论文的特征词空间向量。

2.2.4 Stacking 模型分类模块

Stacking 模型分类模块是整个系统的核心,具体运行中采用 10 折交叉验证。

(1)在训练子系统中,以特征集与第一个类别作为输入数据,首先,以 10 折交叉验证的方法对四个基分类器进行训练和预测;随后,将各基分类器预测所得结果作为特征项,结合原始数据的第一个类别对元分类器进行训练;最后,将所有训练所得模型保存至模型库。

(2)对剩余所有类别,依次做步骤(1)中操作。

(3)当对所有类别的训练全部完成后,进入预测子系统。以只有特征项没有类别的待预测数据作为输入,四个基分类器分别调取训练第一个类别时生成的模型对其进行预测。随后用元分类器调取训练第一个类别时生成的模型,以四个基模型产生的预测数据作为输入,输出对第一个类别的预测结果。

(4)对剩余所有类别依次重复步骤(3)中操作后,汇总对所有类别的预测结果,输出总的分类结果。

3 系统实现

3.1 实验设计

为验证本研究所构建的基于 Stacking 模型的

学术论文多标签分类系统在理论上的优越性，设计以下仿真实验。从中国知网数据库中随机搜集1200篇在T(工业技术)、V(航空航天)、Q(生物科学)、K(历史地理)四个大类范围内的学术论文作为实验数据，其中T(工业技术)与其他三个大类有着广泛的交叉研究内容。基于python语言进行编程，将本研究设计的学术论文多标签分类系统予以完整实现。

3.2 分类结果

利用基于Stacking模型的学术论文多标签分类系统对搜集到的1200篇论文进行自动分类实验，部分分类结果如图1所示。分类结果可以展示出利用该分类系统对实验数据分类的预测结果及真实结果，并进行对比，检验其结果是否一致。另外，为最大程度保证实验结果的可靠性，避免因偶然误差导致的错误结论，最终采用10折交叉验证法来测试该系统在分类准确率、预测稳定性、泛化能力、抗过拟合能力等方面的性能。

3.3 模型分析

根据学术论文多标签分类的现实需求，本研究将预测准确率、预测稳定性、泛化能力、抗过拟合能力四项指标作为主要评价标准，将Stacking集成分类模型作为评价对象，四个基分类模型(随机森林、支持向量机、极限树、极端梯度提升)为参照对象，构建了评价体系。

	题目	真实类别	预测类别	结果评价
2	公共建筑绿色改造方案设计评价研究	['T']	['T']	正确
3	基于关中传统民居特质的地域性建筑创作模式研究	['T']	['T']	正确
4	无人机在重大地质灾害应急调查中的应用	['V']	['V']	正确
5	基于无人机的建筑物裂缝图像采集与处理系统	['T', 'V']	['T', 'V']	正确
6	面向WSN的无人机水域监测系统研究与应用	['T', 'V']	['T', 'V']	正确
7	多旋翼无人机技术在工程施工管理中的应用	['T', 'V']	['T', 'V']	正确
8	输电线路无人机巡检图像中电力部件识别方法研究	['T', 'V']	['T', 'V']	正确
9	基于无人机移动边缘计算的软件定义网络架构分析	['T', 'V']	['V']	错误
10	应急任务响应时间最优的多星成像规划方法	['V']	['V']	正确
11	高低轨遥感卫星联合监测火灾模式分析	['V']	['V']	正确
12	无人机在消防灭火救援中的应用	['T', 'V']	['T', 'V']	正确
13	警用无人机在边疆地区反恐处突中的应用	['T', 'V']	['V']	错误
14	自然灾害环境中无人应急救援的任务分配研究	['V']	['V']	正确
15	无人机在电力线路巡检中的应用及前景	['T', 'V']	['T', 'V']	正确
16	多旋翼无人机在高层建筑消防灭火中的应用	['T', 'V']	['T', 'V']	正确
17	民用无人机在物流配送行业中的分析与设计	['V']	['V']	正确
18	无人机气象灾害应急平台的设计	['V']	['T', 'V']	错误
19	基于无人机视频分析的道路信息提取技术	['T', 'V']	['T', 'V']	正确
20	输电通道人机协同巡检方式的探索	['T', 'V']	['T', 'V']	正确
21	空天资源对地观测协同任务规划方法	['V']	['V']	正确
22	发展地理学视角下欠发达地区贫困的地方分异与治理	['K']	['K']	正确
23	如何回归地理学:我的思考与实践	['K']	['K']	正确
24	地理学视角的可持续生计研究:现状、问题与领域	['K']	['K']	正确
25	民族旅游村寨居民文化依恋的时空变迁及其机理——以	['K']	['K']	正确
26	地理学碰上“大数据”:热反应与冷思考	['T', 'K']	['K']	错误

图1 Stacking模型分类器的部分分类结果

3.3.1 预测准确率

尽可能提高预测项类别准确程度是分类器的首要目标。多标签分类的准确率会随标签数量的增加而迅速递减，当前文本多标签分类领域很多分类器的预测准确率远未达到理想程度。因此，预测准确率的提升是本研究构建的基于Stacking模型的学术论文多标签分类系统的核心任务。

图2为Stacking模型对Q、T、V、K四个标签二分类准确率和四个标签总体分类准确率的折线图。Stacking模型对Q、V、K三个标签的二分类准确率高达97%，对T标签的二分类准确率略低，约为92%。四个标签总体的分类准确率理论上应为四个标签二分类准确率的乘积，但实际上略有误差，总体分类准确率达到了86%。

图3为Stacking模型与其四个基分类模型在10折交叉验证中十次预测准确率的折线图，由此可知Stacking模型每一次的预测准确率均明显高于其他四个模型。图4为Stacking模型与其四个基分类模型10折交叉验证平均预测准确率的折线图，由此可知其他四个模型平均准确率相差不大，但Stacking模型的平均预测准确率较其四个基分类模型有大幅提升，涨幅达5%以上。

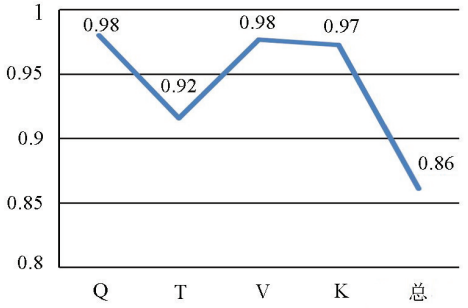

图 2　分类准确率

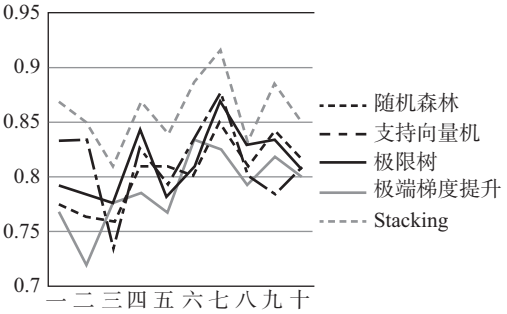

图 3　十次预测准确率

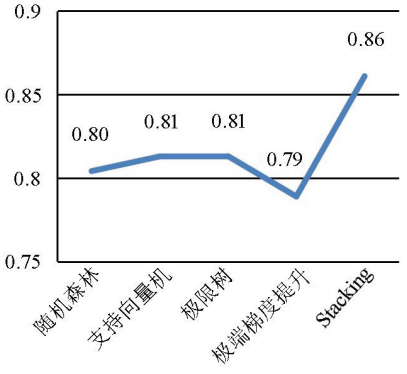

图 4　平均预测准确率

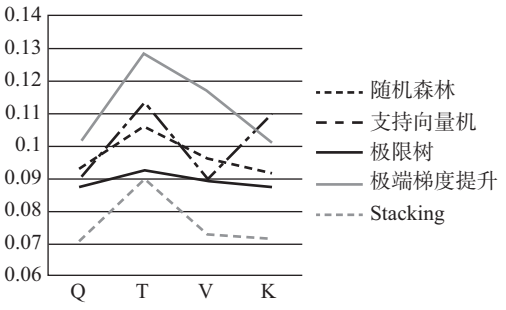

图 5　四个标签的汉明损失

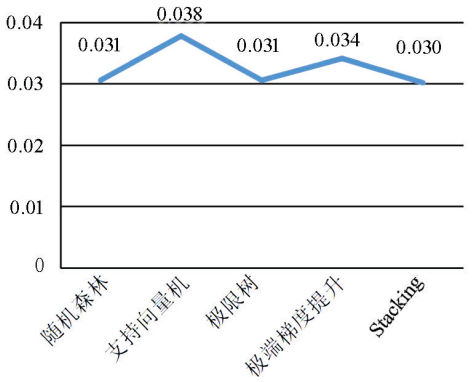

图 6　十次分类准确率的标准差

综上所述，Stacking 模型对于单标签二分类的预测准确率极高，对多标签分类的预测准确率较传统的单模型分类器或同质集成分类器有明显提高。

3.3.2　预测稳定性

在学术论文分类的实际应用中，由于分类数据维度庞大以及模型的不确定性因素较多，分类的准确率会在一定范围内随机浮动。对多二分类框架下的多标签分类而言，分类器判断的次数与标签的数量成正比，而判断次数的成倍增加会导致分类器稳定性降低，从而大幅度增加分类效果不佳的风险。因此，学术论文的多标签分类对于分类模型的预测稳定性要求较高。本研究以多标签分类常用的汉明损失（Hamming Loss）及十次分类准确率的标准差来衡量分类稳定性。

汉明损失可以反映分类错误的标签数目，汉明损失的值越小，说明模型效果越好，分类性能越稳定，其计算如公式（4）。

$$\text{HammingLoss} = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{j=1}^{L}\text{XOR}\left(Y_i^j, P_i^j\right)}{L} \quad \text{公式(4)}$$

图 5 为 Stacking 模型及四个基分类模型的汉明损失比较情况。该图显示，由于标签数据集的不同，各模型在预测不同标签时，其结果稳定性能会产生差异。然而，与其他模型相比，Stacking 模型的汉明损失值在四个标签中均为最低，说明 Stacking 模型的分类错误率最低，稳定性较好。由于模型的稳

定性一定程度上会受到样本的影响,不能仅通过汉明损失指标来衡量,本研究又利用各模型十次分类准确率的标准差进一步分析模型预测稳定性。图 6 为 Stacking 模型及其四个基分类模型在 10 折交叉验证中分类准确率的标准差,以十次分类准确率的标准差代表模型分类的稳定性。由图 6 可知,支持向量机与极端梯度提升预测最不稳定,随机森林和极限树比较稳定,Stacking 模型最稳定,说明 Stacking 模型在分类稳定性方面表现良好。

3.3.3 泛化能力

在文本分类问题中,分类模型针对不同类型的文本其表现往往具有差异性。学术论文的多标签分类问题要求分类模型对不同类型的论文均有着较高的分类准确率,任一标签出现较低准确率都会对总体分类效果产生极大影响。

图 7 为 Stacking 模型及其四个基分类模型在 10 折交叉验证中对 Q、T、V、K 四个标签的平均预测准确率。由该图可知,Stacking 模型在 Q、T、V、K 四个标签上的平均预测准确率都高于其他四个模型。即使在预测准确率最低的 T 标签上,Stacking 模型的预测准确率依然达到了 91.6%,表明 Stacking 模型具有良好的普适性。

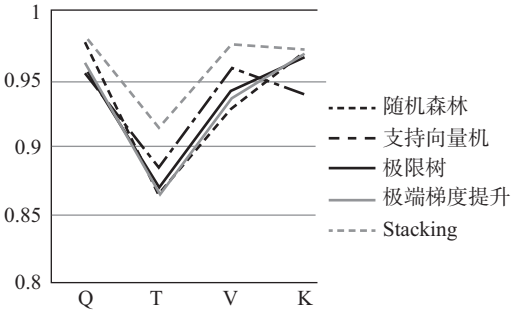

图 7 四个标签的平均预测准确率

图 8 为 Stacking 模型及其四个基分类模型在 10 折交叉验证中对 Q、T、V、K 四个标签预测准确率的标准差,该图显示随机森林、极限树、极端梯度提升三个模型的标准差很大,即在不同类型文本中的发挥稳定性较低。支持向量机模型标准差较小,但仍略大于 Stacking 模型。Stacking 模型的标准差最小,说明对预测不同类型文本的泛化能力较强。

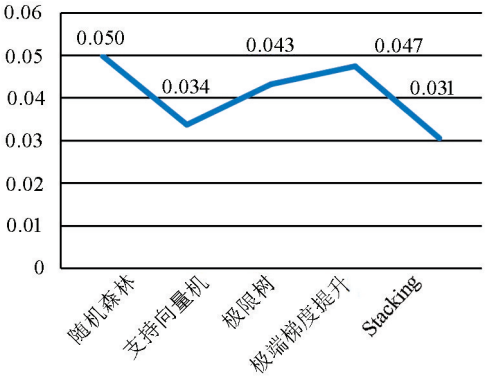

图 8 四个标签预测准确率的标准差

3.3.4 抗过拟合能力

机器学习的众多模型中都存在着不同程度的过拟合现象,通常表现为训练出的模型对训练集的预测准确率很高,但对测试集的预测准确率较低。过拟合现象是制约分类模型准确率提升的重要因素之一,而 Stacking 模型的优势之一就在于有效地克服了过拟合现象。

图 9 为 Stacking 模型及其四个基分类模型在 10 折交叉验证中对训练集和测试集的平均预测准确率。该图显示其他四个模型对训练集的预测准确率大幅高于对测试集的预测准确率,甚至随机森林和极限树模型对训练集的预测准确率已经接近于 1。而 Stacking 模型对训练集的预测准确率基本接近对测试集的预测准确率,表明 Stacking 模型对训练数据的学习程度恰到好处,既没有出现过拟合现象,也未出现欠拟合现象,而其他四个模型均有严重的过拟合现象。

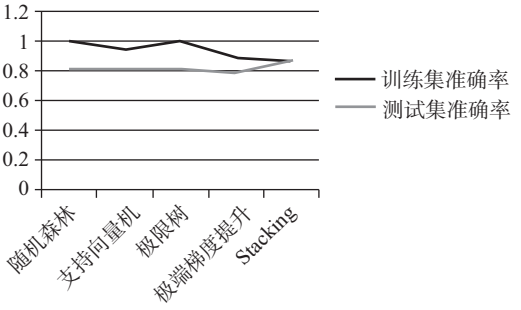

图 9 训练集和测试集的平均预测准确率

综上所述,针对学术论文的多标签分类问题,Stacking模型在分类准确率和抗过拟合能力方面较其他模型有大幅提高。在分类的稳定性和泛化能力方面的表现也较为优秀,有着较高的稳定性和较好的泛化能力。因此,本研究认为Stacking模型在学术论文多标签分类问题上的综合性能显著优于单分类模型和同质集成分类模型。

4 结语

本研究基于Stacking模型、多二分类模型以及多种文本特征表示的技术构建了一个以异质集成分类器为核心的学术论文多标签分类系统。对仿真实验数据的分析,验证了异质集成分类器在解决学术论文多标签分类问题上具有比单分类器或同质集成分类器更高的预测准确率、更好的稳定性、更强的泛化能力,同时较好地克服了过拟合现象。本研究为实现学术论文(尤其是涉及交叉学科研究的论文)高质量分类提供了新的解决思路和实践探索,但所设计的系统仍有一些不足之处。如所使用的仍是较为传统的多二分类模型,忽略了学术论文类别之间的关联性,且并未使用当前机器学习领域最前沿的深度学习技术。后续研究可将Stacking模型与更先进的多标签分类方法或先进的算法相结合,在综合性能上取得更大突破。

参考文献

1 Liakata M, et al. Automatic Recognition of Conceptualization Zones in Scientific Articles and Two Life Science Applications[J]. Bioinformatics, 2012, 28(7): 991-1000.

2 章成志,等.基于全文内容的学术论文研究方法自动分类研究[J].情报学报,2020(8):852-862.

3 黄学坚,等.基于改进型图神经网络的学术论文分类模型[J].数据分析与知识发现,2022(10):93-102.

4 Amanda Clare, Ross D. King. Knowledge Discovery in Multi-label Phenotype Data[J]. Principles of Data Mining and Knowledge Discovery, 2007(8): 42-53.

5 André Elisseeff, Jason Weston. A kernel method for multi-labelled classification[J]. Advances in Neural Information Processing Systems, 2001(1): 681-687.

6 Schapire R E, Singer Y. BoosTexter: A Boosting-based System for Text Categorization[J]. Machine Learning, 2000(39): 135-168.

7 Min-Ling Zhang, Zhi-Hua Zhou. Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(10): 1338-1351.

8 Min-Ling Zhang, Zhi-Hua Zhou. A k-nearest neighbor based algorithm for multi-label classification[J]. IEEE International Conference on Granular Computing, 2005(2): 718-721.

9 Min-Ling Zhang, Zhi-Hua Zhou. ML-KNN: A lazy learning approach to multi-label learning[J]. Pattern Recognition, 2007, 40(7): 2038-2048.

10 李楚贞,等.复杂文本多标签分类算法的设计与仿真[J].计算机仿真,2022(5):299-303.

11 刘爱琴,马小宁.基于概率主题模型的短文本自动分类系统构建[J].国家图书馆学刊,2020(6):102-112.

12 邵孟良,齐德昱.一种改进的随机森林Boost多标签文本分类算法[J].计算机应用与软件,2022(11):215-221,303.

13 马雨萌,等.融合BERT与多尺度CNN的科技政策内容多标签分类研究[J].情报杂志,2022(11):157-163.

14 史佳琪,张建华.基于多模型融合Stacking集成学习方式的负荷预测方法[J].中国电机工程学报,2019(14):4032-4042.

15 李寿山,黄居仁.基于Stacking组合分类方法的中文情感分类研究[J].中文信息学报,2010(5):56-61.

16 李珩,等.基于Stacking算法的组合分类器及其应用于中文组块分析[J].计算机研究与发展,2005(5):844-848.
17 王彦莹,等.基于文本生成技术的历史古籍事件识别模型构建研究[J].图书情报工作,2023(3):119-130.
18 刘丽帆,等.基于学术文献引文内容的跨学科知识流动研究[J].情报理论与实践,2022(6):24-31,47.
19 胡昊天,等.数字人文视角下的非物质文化遗产文本自动分词及应用研究[J].图书馆杂志,2022(8):76-83.
20 彭秋茹,等.面向新时代的人民日报语料中文分词歧义分析[J].情报科学,2021(11):103-109.
21 吕建新,等.基于词向量语义扩展的网络文本特征选择方法研究[J].情报科学,2019(12):47-51.

(刘爱琴 副教授 山西大学经济与管理学院,郭少鹏 山西大学经济与管理学院物流工程与管理专业2021级硕士研究生)

收稿日期:2023-03-28

中国互联网络信息中心发布第53次《中国互联网络发展状况统计报告》

2024年3月22日,中国互联网络信息中心(CNNIC)发布《第53次中国互联网络发展状况统计报告》,报告显示,截至2023年12月:

• 我国网民规模达10.92亿,互联网普及率达77.5%;其中农村网民规模为3.26亿,城镇网民规模为7.66亿;人均每周上网时长26.1个小时。

• 我国非网民规模为3.17亿,较2022年12月减少2688万人;其中农村地区非网民占比为51.8%,60岁及以上非网民群体占比为39.8%。

• 我国手机网民规模达10.91亿人,网民使用手机上网的比例达99.9%,使用台式电脑、笔记本电脑、电视和平板电脑上网的比例分别为33.9%、30.3%、22.5%和26.6%。

• 我国网站数量为388万个;APP在架数量261万款、小程序超700万个;移动互联网接入流量达3015亿GB,同比增长15.2%。

• 网络视频用户规模为10.67亿,占网民整体的97.7%;短视频用户规模为10.53亿,占网民整体的96.4%。

• 网络直播用户规模达8.16亿,占网民整体的74.7%。

• 网络文学用户规模达5.20亿人,占网民整体的47.6%。

资料来源

第53次中国互联网络发展状况统计报告[R/OL].[2024-04-02].https://www.cnnic.net.cn/NMediaFile/2024/0325/MAIN1711355296414FIQ9XKZV63.pdf.

(国家图书馆研究院 提供)