面向图书主题分类的随机森林算法的应用研究

孙彦雄,李业丽,边玉宁
(北京印刷学院,北京 102600)

摘 要: 针对传统随机森林算法对文本特征提取质量不高导致分类效果差的问题,提出一种对图书等大数据量文本信息文本的改进的随机森林算法。又由于传统随机森林决策树质量难以保证,提出一种加权投票提高决策树质量的机制。算法主要由两方面组成,一方面是基于文本主题特征提取的 Tr-K 方法,目的是提高文本主题特征的质量与代表性;另一方面是基于 bootstrap 抽样时遗留的 1/3 袋外数据提出的验证机制。文中采用的是 20 Newsgroups 数据集和来自于搜狗实验室提供的中文分类语料库,中英文两种数据集充分考虑了该模型的泛化性,并在实验中验证了不同数据集下较传统随机森林算法拥有更优秀的分类能力。Python 环境下的实验数据表明,该方法在文本分类中相对于 C4.5、KNN、SVM、原始随机森林算法可以取得更好的结果。

关键词: 图书文本分类;随机森林;Tr-K 方法;TRk-SW-RF 模型;主题分类;决策树
中图分类号: TP301.6 **文献标识码:** A **文章编号:** 1673-629X(2020)06-0065-06
doi:10.3969/j.issn.1673-629X.2020.06.013

Application of Random Forest Algorithm for Book Subject Classification

SUN Yan-xiong,LI Ye-li,BIAN Yu-ning
(Beijing Institute of Graphic Communication,Beijing 102600,China)

Abstract: In view of the problem of poor classification effect caused by low quality of extracting text features for the traditional random forest algorithm,an improved random forest algorithm for the text of big data like books is proposed. Since the quality of traditional random forest decision tree is difficult to guarantee,a weighted voting mechanism to improve the quality of decision-making tree is presented. The algorithm is mainly composed of two aspects. One is the Tr-K method based on text theme feature extraction,which aims to improve the quality and representation of text features. The other is the verification mechanism of 1/3 of the extra-bags of data left over from the bootstrap sampling. We use the 20 Newsgroups dataset and the Chinese corpus from the Sogou Lab. For the Chinese and English datasets,we take full consideration of the generalization of the model and verify that it has better classification ability compared with the traditional random forests under different datasets. The experimental data in Python environment show that the proposed method can achieve better results in text classification relative to C4.5,KNN,SVM and original random forest algorithm.

Key words: book text classification;random forest;Tr-K method;TRk-SW-RF model;theme classification;decision tree

0 引 言

随着科技的快速发展,自从 20 世纪 90 年代以来,产生的数据信息越来越多,其中 80%的信息是以文本存储的[1]。因此,人们对巨量的文本信息,不能采用之前传统的人工筛选。基于自然语言处理的文本处理应运而生,近年来对文本分类的研究也越来越多,主要集中在朴素贝叶斯、K-means 聚类、SVM 等算法。随机森林算法(random forest,RF)由于其具有训练速度快、对于大数据时代易于进行并行计算、具有很强的抗干扰能力且抗过拟合能力优秀的优点,被应用在各行各业中,并取得了优于传统方法的效果。

对于文本分类的研究,已有大量的研究成果。比如,周庆平提出了一种基于聚类的改进 KNN 算法[2];Yang 对特征选择函数进行了改进,将几种特征选择函数的准确系数连接起来构成一种新的特征选择函数,最后再使用 SVM 进行分类[3];张翔提出了使用 Bagging 的中文文本分类器的改进算法[4]。基于文本信息的增加和文本处理技术的发展,对于文本分类的应用也越来越多。例如舆情监测、情感分析、商品分类、新闻分类等等。

由于随机森林算法的诸多优势,使得专家学者对 RF 进行了许多改进应用研究。1995 年 Tinkam-ho 首

收稿日期: 2019-07-10 **修回日期:** 2019-11-12
基金项目: 北京市科技创新服务能力协同创新项目(PXM2016_014223_000025)

作者简介: 孙彦雄(1995-),男,硕士,研究方向为数据挖掘、自然语言处理;李业丽,教授,研究方向为出版大数据、数据挖掘、信息处理技术。

次提出随机森林的概念[5]。之后 Leo Boeiman 提出 RF 是一种分类和预测模型[6]。M P Perrone 等人提出在分类阶段,RF 类标签是由所有决策树的分类结果综合而成,并在投票[7]跟概率平均[8]两个方面是使用最多的方法。在应用方面,在生物信息学中,EI-Atta 提出了一种使用 RF 预测大麻素受体(CB2)激动剂活性的方法[9];生态学中,Eruan 等人使用 RF 对空气预测进行了研究;在遗传学中,Retralia 在基因识别上使用了 RF。并且,RF 在生物芯片、信息抽取等领域,均取得了不错的效果。

1 算法介绍

随机森林算法是由许多决策树构成,通过每个决策树的决策结果进行投票,获得票数最多的类别就是随机森林算法的结果。由于随机森林算法的可并行计算[10],容易泛化应用,不易过拟合等优点,其在生物[11]、医学[12]、信息检索等多个领域得到了广泛的应用。文中主要内容是讲解 RF 的基本构建流程,决策树的基础知识,为后续对 RF 的优化做好铺垫工作。

1.1 随机森林算法

要了解随机森林算法,首先就要明白决策树的由来。因为 C4.5 是决策树的经典算法结构,因此首先分析解释 C4.5。虽然决策树有很多变种,但是其核心主函数类似,不同点在于最优特征标准的选择上。

首先了解信息熵,信息熵是香农在 1948 年提出的概念,用来解决信息的量化度量问题,如式(1)所示。

$$E(X) = -\sum_{i=1}^{r} P(u_i)\log P(u_i) \tag{1}$$

其中,i 为每个消息,r 为消息的个数。

为了便于后面的计算,文中提出将累加符号前面的负号删掉,定义为纯度。相应的信息熵越低,对应数据集的纯度越高。

$$\mathrm{Ent}(D) = \sum_{k=1}^{y} P_k \log_2 P_k \tag{2}$$

式(2)就是文中定义的纯度计算公式,D 表示某个数据集,k 表示由于某种属性导致的分类,总共分为 y 类。假设由于属性 t 导致数据集 U 分为若干类,由此计算属性 t 的信息增益,如下式:

$$\mathrm{Gain}(D,t) = \mathrm{Ent}(D) - \sum_{U=1}^{U} \frac{|D^U|}{|D|}\mathrm{Ent}(D^U) \tag{3}$$

由此,可以使用式(3)作为最优特征标准,得到的是 ID3 算法,但是由于 ID3 算法不能处理连续型的变量属性,并且在属性偏好上偏向于属性分类多的属性,影响决策。故而,1993 年 Quinlan 提出了 C4.5 算法,使用信息增益率作为最优特征标准,从而解决了 ID3 的属性偏好等问题。信息增益率计算公式如下:

$$\mathrm{Gain\_rotio}(D,t) = \frac{\mathrm{Gain}(D,t)}{\mathrm{IV}(t)} \tag{4}$$

$$\mathrm{IV}(t) = -\sum_{U=1}^{U} \frac{|D^U|}{|D|}\log_2 \frac{|D^U|}{|D|} \tag{5}$$

C4.5 算法的算法步骤如下:

(1)对数据集进行预处理,将连续变量离散化或对缺失的数据进行补充,进行预处理验证结束之后,进行下一步;

(2)判断本训练集中是否已经生成叶子节点,如果已经全部生成,结束算法,进行下一步;

(3)使用式(4)计算叶子节点的信息增益率,进行下一步;

(4)比较得出使信息增益率最大的属性,作为分裂节点,再使用所选属性分割训练集为一个个的子训练集,转为第二步。

由此,可以得到一棵棵的决策树,随机森林算法利用多棵单决策分类器组合而成多分类器的思想,克服了单分类器决策树的种种缺点。通过每棵决策树的结果进行投票,得到最终的分类结果。这就是 RF 的主要思路,构建过程是,首先利用 bootstrap 对训练集进行抽样生成多个新的训练集;其次利用生成的训练集产生一棵棵决策树;最后通过每棵决策树的投票得出最终的分类结果。

RF 的主要关键点在于两次随机抽样,一次是使用 bootstrap 有放回抽样,会使得新的数据集中包含 2/3 旧数据集的内容,从而产生袋外数据,为文中的改进优化提供了数据源。另一次是随机抽样,产生在对特征的选择上,每次生成决策树时,使用的特征并不是完全相同的。从给出的特征中随机抽取少于总特征数的特征进行决策树的生成。将前面生成的决策树,一棵棵的连起来就成为了随机森林。

1.2 RF 算法步骤

随机森林算法在文本分类中的算法步骤如下:

(1)文本预处理。首先去除文本中的停用词、符号等“噪声”。然后使用 word2vec 词嵌入模型,对文本信息进行向量化,生成训练集。

(2)假设训练集中包含有 N 个样本,T 种分类属性。采用 bootstrap 抽样方法,抽取出样本 N 个,得到新的样本集。

(3)在给出的 T 种分类属性中,随机抽取 t($t \leqslant T$)种属性。使用某种决策树最优特征标准,选择最优分类节点,使得在子样本集中均为叶子节点。

(4)重复进行 K 次第三步,生成决策树 K 棵,得到最终的随机森林。

(5) $H(x)$ 是分类器的函数模型,决策树用 h_i 表示,Y 表示目标变量(分类标签),$I(*)$ 表示函数。随

机森林的决策公式如式(6) 。

$$H(x) = \operatorname{argmax} \sum_{i=1}^{k} I(h_i(x) = Y) \tag{6}$$

传统随机森林流程如图 1 所示。

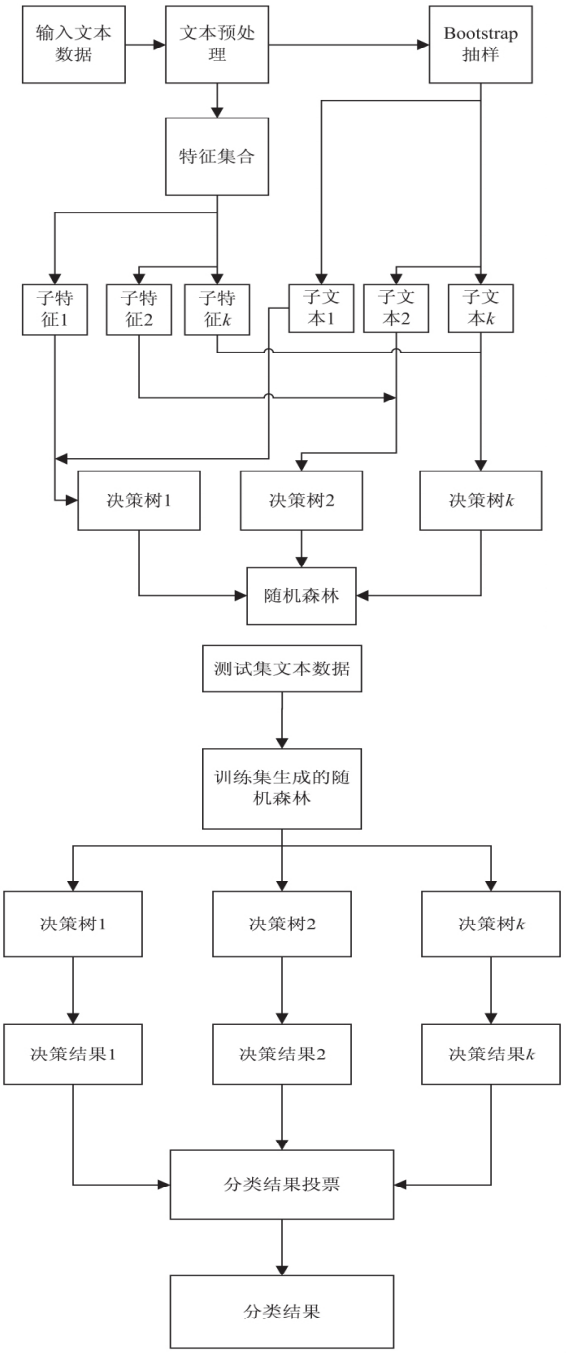

图 1　随机森林算法流程

1.3　RF 优缺点分析

随机森林算法采用的是 bootstrap 抽样，对数据集进行有放回抽样生成子数据集。正是如此，会产生大约 1/3 的袋外数据，这些袋外数据对分析随机森林算法的性能具有很高的研究价值。Breiman 指出，袋外数据可以替换数据集的交叉验证法，并且袋外数据的估计就是 RF 泛化误差的无偏估计。文中正是利用袋外数据，对 RF 的决策树在相似程度高的情况下会掩盖真实分类的缺点进行了改进。

随机森林算法是一种集成学习算法，将多个弱分类器组合使用，得到一个分类性能更强的强分类器。RF 正是因为分开生成决策树，所以便于并行化处理数据[13-14]。在大数据时代，可以并行化处理数据的优势，使得该算法极具诱惑力。

2　改进的 RF 算法

首先传统的随机森林算法没有考虑文本数据的特殊性，进而在处理文本数据时，往往由于特征提取质量差，不能提高文本分类的水平[15-16]。其次，随机森林算法本身存在着相似决策树会掩盖真实分类决策的问题。文中针对以上两个方面进行改进。

2.1　提高文本特征质量的 Tr-K 方法

传统的随机森林算法在进行分类决策时，特征选择个数、质量的问题并不突出。但是对于图书等大容量的文本进行分类的问题来说，文本特征(分类决策树属性) 数量越多、质量越高，得到的分类效果就会越好。为此，文中提出一种 TF-IDF、TextRank、K-means 三种方式结合的 Tr-K 方法，以提高文本分类效果。

TF-IDF 方法的全称是 term frequency-inverse document frequency，是用来进行信息检索和数据挖掘的常用技术。TF 值是指在文件 j 中，第 i 个词的重要程度。

$$\mathrm{TF}_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{7}$$

其中，分子表示第 i 个词在文件 j 中的出现频数，分母表示文件 j 中包含有 k 个单词出现的总频数。

$$\mathrm{IDF}_i = \log \frac{|D|}{|\{j: t_i \in d_j\}| + 1} \tag{8}$$

其中，$|D|$ 表示语料库中的文件总数，t_i 表示要检验的第 i 个词，d_j 表示文件 j 包含的词汇集合，$|\{j: t_i \in d_j\}|$ 表示包含 t_i 的所有文件频数。之所以在分母需要加 1 操作，是为了避免出现无意义的除零情况的发生。

$$\mathrm{TFIDF}_{ij} = \mathrm{TF}_{ij} \times \mathrm{IDF}_{ij} \tag{9}$$

通过式(8) 、式(9) ，对某一文件内的高词语频率和在文件集中的低文件频率，产生权重高的 TF-IDF。从结论可以看出，该算法倾向于过滤掉常见的词语，保留重要的词语。缺点是，文本的开头跟结尾对于语义具有不同的重要性，不能体现词语的位置信息。

TextRank 算法来源于 Google 的 PageRank 算法，PageRank 算法是用来评判一个网页的重要程度，采用有向无权图进行打分[17-18]。设定 V_j 表示网页 j 的节点，$\mathrm{In}(V_j)$ 表示指向网页 j 的节点集合，$\mathrm{Out}(V_j)$ 表示

网页 j 指向其他网页节点的集合。$|\mathrm{In}(V_j)|$ 表示集合的节点数量。对网页重要性的打分如式(10)。

$$S(V_i) = (1-d) + d \times \sum_{j \in \mathrm{In}(V_j)} \frac{S(V_j)}{|\mathrm{Out}(V_j)|} \tag{10}$$

d 表示阻尼系数,当极端情况下没有人访问的网页,会使公式无意义,因此加入阻尼系数,避免此情况的发生。

TextRank 是针对文本关键词的提取方法,相对于网页,最大的不同在于生成的是有向有权图。使用一个可变窗口对文章句子进行扫描,去掉停用词之后,认定窗口内的每个词之间都是互相联系的,并计算相邻词语之间的余弦相似性,计算每个词汇之间的权重。由此,计算公式如下:

$$\mathrm{WS}(V_i) = (1-d) + d \times \sum_{V_j \in \mathrm{In}(V_j)} \frac{W_{ji}}{\sum_{V_k \in \mathrm{Out}(V_j)} W_{jk}} \times \mathrm{WS}(V_j) \tag{11}$$

其中,$\mathrm{WS}(V_i)$ 表示节点 V_i 的得分,W_{ji} 是指用余弦相似度计算的节点 V_j 对 V_i 的权重。此种关键词的提取方法的优点在于计算复杂度小,容易处理。

K-mean 算法是无监督分类算法,因为其计算简单聚类效果优,故而应用广泛。但是使用 K-means 算法,关键在于找到合适的 k 值,即初始化质心。选择的恰到好处,会明显提高算法的效率和准确度。为此,文中采用 TF-IDF 与 TextRank 提取出的特征集作为 K-means 算法的 k 值输入,再进行聚类分析。综上所述,提出的 Tr-K 文本特征抽取模块的流程如图 2 所示。

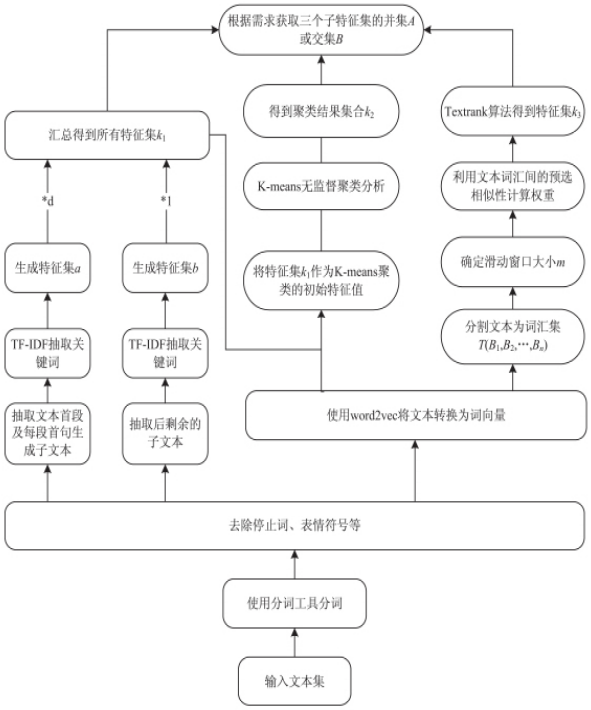

图 2 特征提取 Tr-K 方法的结构

2.2 加权投票提高决策树质量的机制

文中使用 Tr-K 得到子特征集,采用 C4.5 算法得到决策树节点集合。基于投票与余弦相似性原理提出优化随机森林机制,改进内容主要包括如下两点:一方面,对训练集通过 word2vec 训练为词向量,再将得到的决策树节点集合转换为词向量,通过使用余弦相似性公式,判定出各个决策树的相似性;另一方面,由于随机森林算法在对训练文本集进行 bootstrap 抽样时,会遗留 1/3 的袋外数据,故而使用袋外数据对决策树节点集合生成的决策树进行准确率测试。

相似度的计算公式如式(12):

$$\mathrm{Similarity} = \frac{\boldsymbol{A} \bullet \boldsymbol{B}}{\|\boldsymbol{A}\| \|\boldsymbol{B}\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{12}$$

其中,A 与 B 分别表示两个节点集合词向量组合,n 表示词向量的维数。通过计算各个节点集合的相似性,选择最优参数 d 进行比较,抽取出相似性较小的节点集合作为备选决策树集合。

当使用训练集 bootstrap 抽取子文本集合时,随之生成袋外数据,因此使用这部分袋外数据,对决策树进行准确率测试,从而找出最优的决策树作为备选。公式如下:

$$\mathrm{Weight}(i) = \frac{N^{\mathrm{cor}}(i)}{N},\ e = \frac{\sum_{i=1}^{k} \mathrm{Weight}(i)}{k} \tag{13}$$

其中,N 是总体样本数(袋外样本数),N^{cor} 是正确分类数,$\mathrm{Weight}(i)$ 表示决策树 i 的正确率,e 是用于判定决策树优劣的阈值。

实现流程如图 3 所示。

3 实验仿真及分析

3.1 实验环境

算法性能分析的实验环境以 Windows 7 操作系统作为整个实验的支撑,使用 Python 语言作为程序设计的编译语言。相关配置:Intel(R) Core(TM) i5-3470 CPU 处理器,处理器为 3.20 GHz,内存为 8 G,开发工具为 PyCharm Community Edition 2017.5.1。

3.2 实验数据集

文中共使用了两种数据集,为验证文中算法在不同文本中的分类能力,特选取中英文两种不同的数据集。20 Newsgroups 数据集共有 18 000 篇文章,涉及到 20 种主题,分为训练集和测试集,两者无交叉。通常是用来进行文本分类、信息检索和文本挖掘相关研究

的国际通用标准数据集。20 Newsgroups 数据集文档分布如图 4 所示。

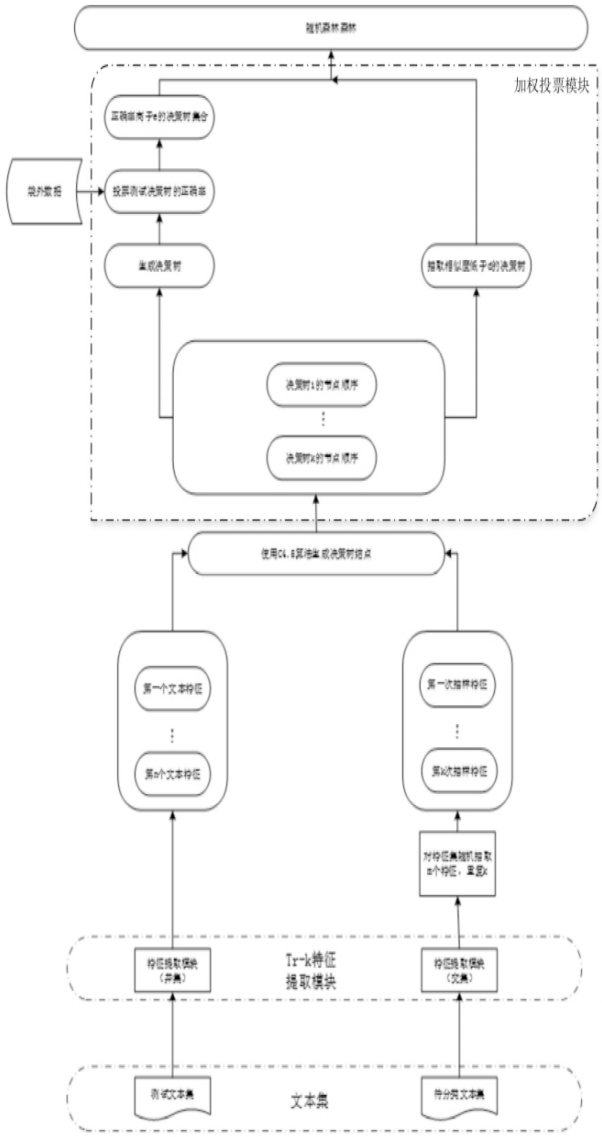

图 3　加权投票模块

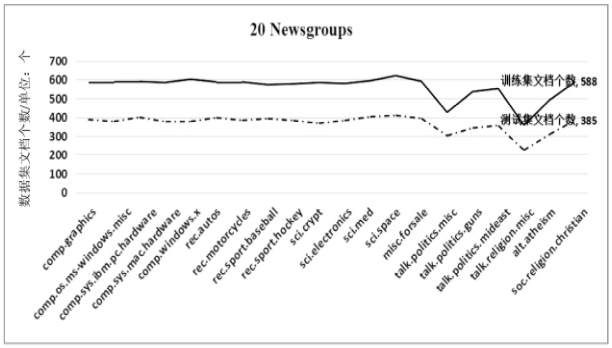

图 4　20 Newsgroups 数据集分布

中文数据集来自搜狗实验室提供的中文分类语料库，随机选取部分语料作为训练集跟测试集。中文数据集包含体育、财经、娱乐等八类新闻文档，共计四万篇。各个类别的数量分布较均匀，大约每类 5 000 篇文档。图 4 横坐标表示 20 个主题的分类名称，纵坐标表示每类主题所包含的文档个数。

3.3　实验设计与分析

利用传统随机森林算法、仅采用 TF-IDF 进行文本特征提取的 TF-RF 算法，与文中设计的 TRk-SW-RF 模型进行对比实验分析。实验对比主要分为以下几个方面: 运行时间、分类准确率和 F_1 值。并且为保证实验数据的稳定性，在决策树数量分别是 50、70、100、200、300、400 时进行对比实验，并且在实验环境不变的前提下运行 10 次，取平均值作为最终的实验结果。

如表 1 所示，当控制决策树的个数时，记录 20 Newsgroups 数据集在不同模型上的训练时长，可以看到，虽然 TRk-SW-RF 模型的训练时长跟传统随机森林的训练时长相差不大，虽然增加了特征选择跟投票的时间，但是由于缩小了文本长度从而使得总的训练时长增长并不明显。并且当决策树的数量增长到一定量时，训练时长的增长速度明显小于决策树的增长速度。

表 1　20 Newsgroups 数据集下各模型运行时长

决策树个数	训练时长/s		
	传统随机森林算法(RF)	IDF-RF	TRk-SW-RF
50	17.12	17.55	17.67
100	32.61	34.01	34.74
200	67.24	68.37	69.11
300	100.03	103.19	104.66
400	133.79	135.94	136.78

同时计算了 20 Newsgroups 数据集在三种有效模型下的准确率，如图 5 所示。从图上可以清楚地看到，使用 TRk-SW-RF 模型的准确率明显高于传统随机森林算法和 IDF-RF 算法。

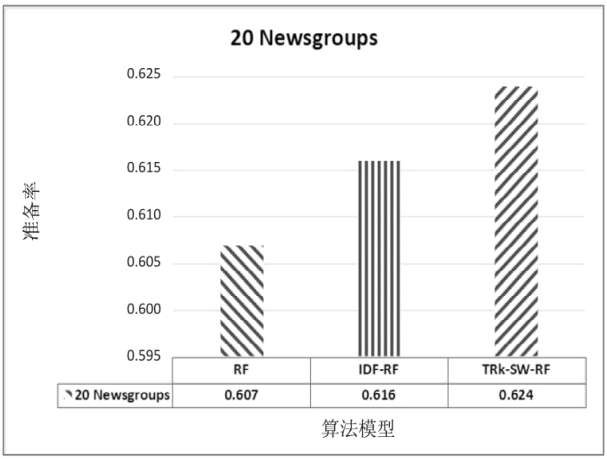

图 5　不同模型下的文本分类准确率

同时，为丰富文中模型在不同数据集上的泛化性，对搜狗实验室提供的中文文本分类语料库，使用 F_1 值的评价标准对三种模型进行评测。结果如图 6 所示，可以看出在中文语料库中文中模型的分类能力仍然高于其他两种算法。

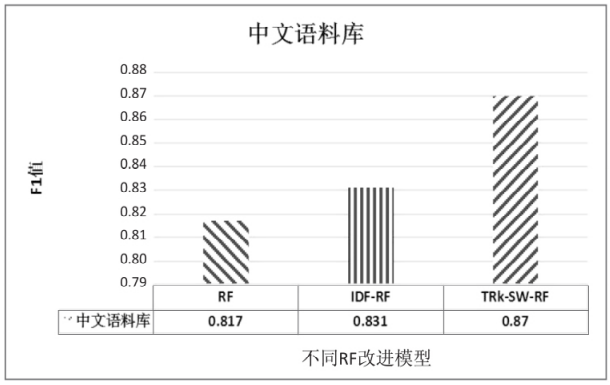

图 6 不同 RF 模型在中文数据库下的 F_1 值

4 结束语

通过对随机森林算法的输入文本数据集进行处理，从而提高分类效果。并使用剩余的袋外数据，对得到的决策树做进一步的检测提取，从而提高了决策树的质量，进而提高了最终生成的随机森林的分类准确率。通过中外两种不同的数据集，验证了该模型在没有明显提高训练时间的前提下，有效地提高了分类准确率和 F_1 值。

下一步的研究可以在计算文本相似度方面引入目前很热门的深度学习算法，虽然会延长训练时间，但是从传统机器学习的分类效果上看，如果能够有较大的提高，还是很有必要的。

参考文献:

[1] KORDE V.Text classification and classifiers: a survey [J].International Journal of Artificial Intelligence & Applications,2012,3(2):85-99.

[2] 周庆平,谭长庚,王宏君,等.基于聚类改进的 KNN 文本分类算法 [J].计算机应用研究,2016,33(11):3374-3377.

[3] KIBRIYA A M,FRANK E,PFAHRINGER B,et al.Multinomial naive Bayes for text categorization revisited [C]//Australasian joint conference on artificial intelligence.Berlin: Springer,2004:488-499.

[4] 张 翔,周明全,耿国华.Bagging 中文文本分类器的改进方法研究 [J].小型微型计算机系统,2010,31(2):281-284.

[5] HO T K.Random decision forest [C]//Proceedings of the 3rd international conference on document analysis and recognition.Montreal,Canada: [s.n.],1995:278-282.

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

[7] HANSEN L K,SALAMON P.Neural network ensembles [J].IEEE Transactions on Pattern Analysis and Machine Intelligence,1990,12(10):993-1001.

[8] PERRONE M P,COOPER L N.When networks disagree: ensemble method for neural net works [C]//Artificial neural networks for speech and vision.New York: Chapman & Hall,1993:126-142.

[9] EL-ATTA A H A,MOUSSA M I,HASSANIEN A E.Predicting biological activity of 2,4,6-trisubstituted 1,3,5-triazines using random forest [C]//Proceedings of the fifth international conference on innovations in bio-inspired computing and applications. [s.l.]: Springer,2014:101-110.

[10] SCORNET E,BIAU G,PHILIPPE-VERT J.Consistency of random forests [J].Annals of Statistics,2015,43(4):1716-1741.

[11] PETRALIA F,WANG P,YANG J,et al.Integrative random forest for gene regulatory network inference [J].Bioinformatics,2015,31(12):i197-i205.

[12] MIGUEL L,GIANLUCA B.Experimental assessment of static and dynamic algorithms for gene regulation inference from time series expression data [J].Frontiers in Genetics,2013,4:303.

[13] 张 鑫.随机森林算法的优化研究及在文本并行分类上的应用 [D].南京: 南京邮电大学,2018.

[14] 罗元帅.基于随机森林和 Spark 的并行文本分类算法研究 [D].成都: 西南交通大学,2016.

[15] 刘 勇,兴艳云.基于改进随机森林算法的文本分类研究与应用 [J].计算机系统应用,2019,28(5):220-225.

[16] 刘耀杰, 刘独玉.基于不平衡数据集的改进随机森林算法研究 [J].计算机技术与发展,2019, 29(6):100-104.

[17] 张莉婧,曾庆涛,李业丽,等.面向图书主题的爬虫算法研究 [J].计算机科学,2017,44(11A):460-463.

[18] 王奕森,夏树涛.集成学习之随机森林算法综述 [J].信息通信技术,2018(1):49-55.