doi:10.6041/j.issn.1000-1298.2019.06.032

基于 Spark 框架 XGBoost 的林业文本并行分类方法研究

崔晓晖　师栋瑜　陈志泊　许　福
(北京林业大学信息学院, 北京 100083)

摘要: 针对当前“互联网 +”技术与林业的交叉融合,涌现出海量待挖掘的涉林文本,而林业文本分类的相关研究尚不成熟的问题,使用网络爬虫技术面向互联网采集涉林文本,基于丰富的语料重新构建分类标签,提出基于 Spark 计算框架的 XGBoost 并行化方法,对林业文本进行分类。经由交叉验证,构建的 XGBoost 并行分类算法准确率为 0.923 4,在各类别中最低 F_1 为 0.860 4,最高为 0.998 4;其在 2.1 万条、4.2 万条、8.4 万条数据集上的训练加速比分别为 2.13、3.47、3.82。结果表明,基于该标签设定的分类模型对现存互联网中涉林文本的适应性较好;Spark 环境下实现的 XGBoost 并行化算法的准确率显著优于其他 4 种机器学习(朴素贝叶斯、GBDT 决策树、BP 神经网络和 ELM 神经网络算法)的并行化算法,算法执行效率远高于单机版本,且数据量越大,其加速比越高,能有效应对海量林业文本的实时、准确分类。

关键词: 林业文本; 文本分类; 大数据分析; Spark; XGBoost
中图分类号: TP391.1　　**文献标识码**: A　　**文章编号**: 1000-1298(2019)06-0280-08

Parallel Forestry Text Classification Technology Based on XGBoost in Spark Framework

CUI Xiaohui　SHI Dongyu　CHEN Zhibo　XU Fu
(*College of Information Science and Technology*, *Beijing Forestry University*, *Beijing* 100083, *China*)

Abstract: At present, the cross-integration of computer technology and forestry field had emerged a large number of forestry texts to be explored, and the shortcomings of related research could be summarized in two aspects: the classification labels in the existing classification system were set unscientific, leading to the classification model lacking of ability to classify the texts on net; the classification algorithm was mostly trained in the single-machine environment without considering its parallelism, then the algorithm could not deal with the actual large-scale data classification problem. Therefore, it was pretty realistic and urgency to design more scientific classification labels and classify forestry texts based on Spark framework. A new crawler technology was used to collect forestry-related texts, and re-construct labels by referring to the existing information retrieval system of forestry to improve the adaptability of classification models. Then the XGBoost parallelization implementation method was realized based on Spark, which completed the computing of training and prediction by RDD program mode. Through cross-validation method, the accuracy of XGBoost parallel algorithm could reach 0.923 4. The lowest F1-measure value was 0.860 4 and the highest was 0.998 4. By training on the 21 thousand, 42 thousand and 84 thousand data sets, the speedup ratios could reach 2.13, 3.47 and 3.82, respectively. The results showed that the new classification labels were set more scientific, and the system had better adaptability to the forestry-related texts on the existing internet. The precision and recall values of the XGBoost algorithm were significantly better than the four kinds of parallel algorithms based on Spark which included NB, gradient boosting decision tree, back propagation neural network, extreme learning machine and ran more effective than the stand-alone version. And with the increase of the data number, the acceleration ratio could be improved, which meant it was pretty useful to deal with the problem about the real-time and accurate classification of massive forestry texts.

Key words: forestry text; text classification; big data analysis; Spark; XGBoost

收稿日期: 2019-03-02　修回日期: 2019-04-15
基金项目: 国家自然科学基金项目(61772078)和北京林业大学热点追踪项目(2018BLRD18)
作者简介: 崔晓晖(1984—),男,讲师,博士,主要从事领域大数据分析及智能信息处理技术研究,E-mail: cuixiaohui@bjfu.edu.cn
通信作者: 陈志泊(1967—),男,教授,博士生导师,主要从事大数据技术、计算机软件与理论研究,E-mail: zhibo@bjfu.edu.cn

0　引言

信息资源的开发利用是国家信息化的核心，随着我国信息化建设的深入，物联网、大数据等技术与林业快速融合，大量涉林的信息网站、数据库、林业监测与评估系统等涌现，从而积累了丰富的林业文本信息，亟待挖掘。但是，各系统间的信息缺乏规划与共享，导致林业文本的信息整合水平不足、利用率低下[1]。因此，对互联网中海量林业文本自动进行精准、快速的分类将有助于推进林业信息化进程，为林业信息的挖掘、热点提取、舆情分析、智能信息推送等研究提供科学的理论与技术支持。

文本分类[2]考虑的首要问题是如何准确判断未知样本的类别，常用于文本分类的算法包括 K 最近邻（K-nearest neighbor，KNN）[3]、朴素贝叶斯（Naive Bayesian，NB）[4]、最大熵[5]、支持向量机（Support vector machine，SVM）[6]、决策树[7]、深度神经网络[8]等。文献[9]将 SVM 算法应用于 Web 农业文本，依据数据采集源将文本分为 8 个产品类别，结果显示 SVM 分类性能优于 NB、决策树等算法。

文本分类算法中的另一个问题是如何提高算法的执行效率，目前较可靠的优化方式是将算法并行化，而基于 Spark 环境实现算法的并行化是较为常见的解决方案[10]。

在林业文本分类领域，文献[11]提出使用 TF－IDF 方法结合差分演化算法对 ELM 极端学习机优化的分类算法，文献[12]则使用高斯混合的分类算法，文献[13]引入 LM 模糊神经网络优化的分类算法，三者以花、树、虫、土壤和水作为分类标签进行实验。实验结果证明，三者使用的算法在其数据集上的表现均较好。但其数据采集不够全面，标签设定不够科学，导致其分类模型无法适用于互联网中现有的林业文本的分类，且算法均在单机环境中实现，未考虑算法的并行性，难以应对大批量数据分类[14]。

由相关文献可知，林业文本分类的相关研究尚不成熟，其亟待解决的问题可概述为两点：① 分类标签设置不科学，其分类体系与林业结合程度低、领域覆盖面不足，无法直接应用于互联网中的涉林文本的分类。②分类算法多在单机环境下实现，缺乏算法并行方面的考虑，不具备应对实际大规模数据场景的能力。为解决上述林业文本分类问题，本文建立一套较为科学、完善的林业文本分类标签，提出一种 Spark 框架下的 XGBoost 算法的并行实现方式，基于该设计构建并行化分类器，衡量不同数据集下该算法的效率和准确率，探索其在海量林业文本分类问题上的有效性。

1　相关技术

1.1　文本预处理流程

预处理是文本分类中最为重要的步骤之一，其处理结果直接影响到后续的分类精度。预处理步骤可概括如下：

（1）采用爬虫技术获取相应的涉林文本，去除异常数据后进行内容解析，使用正则表达式对网页标签进行过滤，建立符合条件的标题与正文。

（2）引入开源工具 ANSJ 包进行中文分词。该分词工具基于 n－Gram＋CRF＋HMM 并使用 Java 实现，分词速度达到 200 万字/s，准确率可达到 96% 以上，适用于当前对分词效果要求较高的各类项目，其分词效果如图 1 所示。

（3）使用停用词集合过滤无用的词汇，构建文本的特征词集合。

干旱区/nz,流域/n,土地/n,资源/n,动态/n,监测/vn,专家系统/gi,方法/n,研究/vn,项目/n,主要/b,新疆/ns,干旱地区/nz,自然环境/l,特点/n,微机/n,干旱区/nz,流域/n,土地/n,资源/n,动态/n,监测/vn,专家系统/gi,干旱区/nz,流域/n,地学/n,空间数据/nz,地学/n,图形/n,遥感/n,数据/n,储/vg,管理/vn,分析/vn,遥感/n,技术/n,计算机/n,图形学/gi,方法/n,图形/n,图象/n,分析/vn,遥感/n,特征/n,信息提取/nz,应用/vn,数学/n,理论/n,系统工程/l,理论/n,灰色/n,系统控制/n,方法/n,模糊数学/n,理论/n,图形/n,图象/n,数据/n,逻辑/n,数学计算/n,研究/vn,因素/n,土地/n,资源/n,发展/vn,影响/vn,作用/n,内在/b,规律性/n,干旱区/nz,国土资源/nz,综合治理/l,区域规划/nz,管理决策/nz,科学依据/l,项目/n,点/q,国内/s,地理信息系统/gi,研究/vn,理论/n,基础/n,软件设计/nz,先进经验/n,遥感技术/nz,资源/n,环境/n,系统/n,技术/n,干旱区/nz,流域/n,土地/n,资源/n,动态/n,监测/vn,侧重于/mq,专家/n,知识/n,计算机/n,数学/n,专家/n,知识/n,推理机/gi,构造方法/gi,领域专家/n,知识库/n,系统/n,方法/n,研究/vn

图 1　林业文本的分词结果

Fig. 1　Word segmentation result of forestry text

1.2　基于 TF－IDF 的特征工程

高维度和高稀疏的向量矩阵给计算机的计算量和学习训练过程增加机器负担，且会影响分类精度，为进一步实现特征矩阵降维，需要对文本特征进行特征选择。

向量空间模型（Vector space model，VSM）[15]是文本分类中最常见的特征标识形式。通过使用这种模型，每篇文档被表示为一组特征向量 $\boldsymbol{D}=\{(w_1,f_1),(w_2,f_2),\cdots,(w_i,f_i),\cdots,(w_n,f_n)\}$，其中 w_i 表示在 $\boldsymbol{D}$ 中出现的特征词，f_i 是特征词 w_i 的权值。其中，i 的取值为 $1,2,\cdots,n$，w_i 经由特征词筛选得到，本文中的 f_i 值将通过经典的词频-逆文件频率算法（Term frequency－inverse document frequency，TF－IDF）[16]进行计算。

TF－IDF 是文本挖掘中常用的加权技术之一，

用于衡量一个字或词在语料库中的重要程度,其计算公式为

$$V_{TF-IDF} = V_{TF} V_{IDF} \tag{1}$$

式中 V_{TF}——特征词在文本中出现的频率

V_{IDF}——特征词的逆向文档频率

1.3 XGBoost 算法原理

XGBoost[17]是基于 Gradient Boosting 算法的一个优化版本,其通过将多个回归树模型集成在一起,形成一个强分类器,具有训练速度快、可并行处理和泛化能力强等优势。

该算法的基本思想[18-19]是选择部分样本和特征生成一个简单模型作为基本分类器,在生成新模型时,学习以前模型的残差,最小化目标函数并生成新模型,此过程重复执行,最终产生由成百上千的线性或树模型,组合为准确率很高的综合模型。它的目标函数 O_{obj} 经过泰勒公式展开后,最终化简为

$$O_{obj} = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda} + \gamma T \tag{2}$$

式中 γ——学习率 λ——正则化参数

T——回归树的叶子数量

G_j——一阶导数 H_j——二阶导数

其中,O_{obj} 的大小依赖于 G_j 和 H_j 的值,O_{obj} 值越小,XGBoost 模型的预测和泛化能力就越强。

1.4 Spark 框架

Apache Spark[20]是 UC Berkeley 开源、类 Hadoop MapReduce 的通用并行计算框架,不同的是,Spark 的 Job 中间结果可以保存在内存中,而不需要读写 HDFS,因此,Spark 是基于内存的并行化计算框架,其执行效率较 Hadoop 快数十倍乃至百倍。Spark 通过基于弹性分布式数据集(Resilient distributed dataset,RDD)的编程模式,使得大部分数据并行算法均可运行于 Spark 集群中。

2 Spark 下林业文本分类算法的并行化

2.1 林业分类标签的设定

文献[21]经调研将林业信息中各个类别的内容具体化,设计出科技、生产资料、市场、花卉、政策等类别,经过对爬虫获取数据的相似性比对,生产资料类与市场类文本存在大量的信息重叠,故将两者合并为林业市场与产业类。根据文献[22]提出林业科技类成果所具有的特点,将科技类报道与技术类成果组合为林业科学与技术类(包含林业论文、林业专利和科学类新闻等)。文献[23]指出林业资源监管中主要包含森林资源、湿地资源以及生物多样性资源等,而花卉植被属植物类,占生物多样性较大比重,因此,将生物多样性资源分类为动物类与植物类。至此,在总结前人研究的基础上,将整体样本分为林业新闻与政策类(A 类)、林业科学与技术类(B 类)、林业市场与产业类(C 类)以及林业资源类(D 类)4 类,并将采集到样本量最多的林业资源类文本分成 4 个子类,即森林类(D1 类)、植物类(D2 类)、动物类(D3 类)、湿地类(D4 类)。相较以往分类体系,该体系分类标签设定更为科学、全面,使得分类模型与林业领域结合更为紧密,也有利于未来更细层面的林业文本分类研究。

根据文献[24]提供的爬虫思路,从互联网中采集原始数据,爬虫语料中约 75% 的文本来源于中国林业新闻网、中国林业政府网、林业信息网、林业产业网等林业相关网站,其余约 25% 来源于新闻刊物,如新华网、绿色时报、百度新闻等综合型新闻网站。随后,从每类中提取 3 000 篇文章,即所有实验样本数为 21 000,将数据按各自标签存入 Hadoop Hive 数据仓库作林业语料。

2.2 基于 Spark 文本分类的并行化设计

基于 Spark 的林业文本分类流程主要分为预处理过程、训练与测试过程。其中,文本预处理中的预处理、特征值计算以及特征词的选取均基于 RDD 并行化实现(图 2)。该预处理程序由 Driver 模块、Mapper 模块以及 Reducer 模块组成。Driver 用于与底层沟通,初始化集群组件;Mapper 模块用于将包含原始文本的 RDD_data 进行去噪、分词形成新的 RDD_words,随后执行 Reducer 模块,基于 CHI 值进行筛选并使用 TD-IDF 进行各个特征词的权重计算,随后生成词向量形式的 RDD_vec。

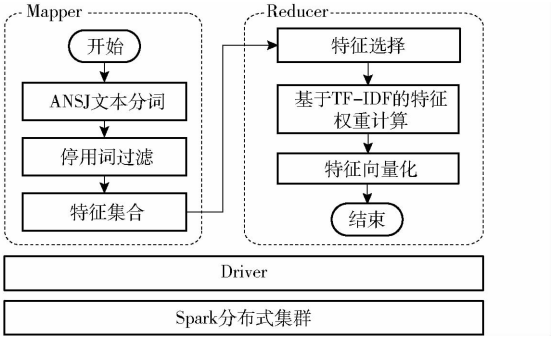

图 2 文本预处理的并行化实现

Fig. 2 Parallelization of text preprocessing

基于 Spark 框架下 XGBoost 的并行化思想是通过 RDD 算子与框架的内存迭代机制提高算法的执行效率来实现的。其中,Spark 集群各节点读取训练数据 RDD_vec 存于内存中。Mapper 部分主要完成决策树的学习过程:在选取分裂节点时,并行计算各个特征的增益,选取增益最大的特征进行分裂以进行树的构建;树的各分支的节点选取都通过并行化进行训练,在达到建树的最大深度或分类增益小于

设定阈值后停止建树，从而完成一次算法中的多个树模型的生成工作；随后由 Reducer 比较并构建准确率相对更高的树模型，输出一轮迭代的结果，随后将迭代结果输入到下次迭代中，直到选取出最优模型。

林业文本的训练与测试过程主要分为如下步骤（图 3）：

（1）对语料中的文本进行自动分词，去除低频词与停用词，构建为〈标签，（文本，特征词集）〉的键值对形式，存入 RDD_data。

（2）通过利用 TF－IDF 进行特征词的权重计算并进行特征向量化，形成〈标签，（文本，特征词集，TF－IDF 权值）〉键值对形式的 RDD_vec。

（3）提取 RDD_vec 中的〈标签，（特征词集，TF－IDF 权值）〉，通过 Spark 提供的转换算子与执行算子构造 XGBoost 与其他 4 种算法的并行分类器。

（4）以随机选取的方式将 90% 的键值对作为训练集 RDD_train 传入分类器，分类器进行迭代训练，并将结果与模型保存。

（5）将余下 10% 的键值对作为测试集 RDD_test 对保存模型的精准率进行验证。

（6）重复步骤（4）～（5），选取最优参数组合，将最优模型保存在 Hive 数据仓库中；基于此，模型将不断进行新数据的训练，从而积累较为科学的林业语料。

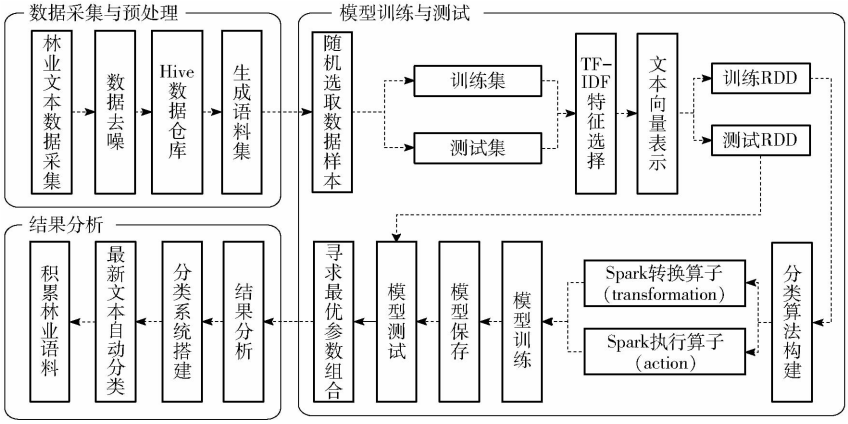

图 3　基于 Spark 的文本分类处理流程图

Fig. 3　Process of forestry text classification based on Spark

3　实验与结果分析

3.1　实验环境与评价指标

采用的硬件环境是 Centos7、Hadoop 2.7.0、Spark 2.2.0、Hive 2.1.1 构成的仿真平台。实验环境共由 5 台主机构成 Spark 计算集群，其中 1 台为 Master 节点，其余 4 台为 Slave 节点，各工作节点的运行内存为 4 GB。

对分类效果的评价采用精准率 P（precision）、召回率 R（recall）、综合评价指标 F_1（F1-measure）、准确率 A（accuracy）等指标，其计算公式为[25]

$$P = \frac{T_P}{T_P + F_P} \tag{3}$$

$$R = \frac{T_P}{T_P + F_N} \tag{4}$$

$$F_1 = \frac{2PR}{P + R} \tag{5}$$

$$A = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{6}$$

式中　T_P——将正类预测为正类的样本数

T_N——将正类预测为负类的样本数

F_P——将负类预测为正类的样本数

F_N——将负类预测为负类的样本数

精准率衡量类别的查准率，召回率衡量类别的查全率，F_1 则综合了 P 和 R 的结果，所以 F_1 越高则说明实验方法越有效，分类器的分类性能越好。

3.2　语料的特征词分析

对预处理后的 21 000 个林业文本语料进行分词统计，共计 4 402 145 个有用词条，无重复词集共计 264 423 个，平均一篇文章中影响分类的词数约为 209 个；篇幅最长的样本包含词数为 11 806 个，最短一篇包含词数为 34 个。现分别为每个类别计算候选词的 TF－IDF 值，并选取每个类别的前 10 个特征词，如表 1 所示。

由表 1 可见，语料中各类别中最为靠前的 10 个特征词频中，仅存在少量交集，且与该标签对应的林业专题动态信息的主题相符，“改革”与“林权”等词汇高频出现与国家推动林业改革的情形相符，说明该分类体系可用于进一步提取林业领域的“热词”；而科学与技术类、市场与产业类以及植物类的高频词汇，一定程度上反映出林业研究多集中于花卉、林木，出现这些词频的样本在结合“花瓣”、“花种”等领域专属名词集合时，即可为挖掘文章主题、提取信息主干等研究提供新思路。因此，本文设计的分类

标签相比原有分类标签更为科学,且有益于林业文本的拓展研究。

表 1 各类别林业文本的前 10 个特征词

Tab. 1 Top 10 characteristic entries of each category

编号	A 类	B 类	C 类	D 类			
				D1 类	D2 类	D3 类	D4 类
1	林业	研究	产业	森林	植物	湿地	野生动物
2	工作	技术	发展	森林资源	花卉	生态	记者
3	改革	建设	生产	林地	生态	自然保护区	动物
4	发展	花卉	试验	面积	苗木	生态系统	资源
5	林权	工作	方法	木材	产品	生物	鸟类
6	生态	种苗	成果	防火	绿色	多样性	湿地
7	森林	林木	项目	林木	景观	湖泊	中国
8	林地	经济	树种	调查	花种	野外	保护区
9	建设	旅游	发展	监测	研究	区域	数量
10	政策	质量	系统	生态	树种	管理	物种

3.3 各并行算法评价指标的对比

使用前文中叙述的分类实验方法,从 Hive 数据仓库中,每类调取 3 000 个样本作为实验数据,随机选取其中 90% 样本为训练集,其余 10% 的样本为测试集。为验证并行化环境下,XGBoost 与传统机器学习算法、基于神经网络的算法的性能,实验选取了 NB、GBDT 决策树代表传统机器学习,选用 ELM 极端学习机和 BP 神经网络(Back propagation neural network,BPNN)作为神经网络算法的代表,将经过网格搜索获取其最优参数的 5 种并行化算法测试结果的 F_1 记录为表 2。

由表 2 可知,在 100% 的数据集下,XGBoost 在每个类别上的 F_1 均高于其他 4 种机器学习算法。模型训练过程与 XGBoost 相似的 GBDT 的性能仅次于 XGBoost 算法。在面对海量文本的分类场景中,两种基于神经网络的算法 ELM 与 BPNN 整体分类效果不如其他 3 种算法。

XGBoost 算法在 A、C、D4 这 3 类上的 F_1 分别为 0.998 4、0.982 9 和 0.945 6,尤其 A 类与 C 类的 F_1 几乎达到 1.0,说明林业新闻与政策类和林业市场

表 2 各文本分类算法的 F_1 对比

Tab. 2 Comparison of F_1 values of each text classification parallel algorithm

类别	XGBoost	NB	GBDT	ELM	BPNN
A 类	0.998 4	0.826 4	0.904 9	0.672 8	0.732 7
B 类	0.896 3	0.756 4	0.719 5	0.647 8	0.736 8
C 类	0.982 9	0.739 5	0.920 1	0.709 7	0.728 1
D1 类	0.890 0	0.756 6	0.661 8	0.616 2	0.683 9
D2 类	0.887 2	0.860 9	0.803 8	0.836 0	0.821 7
D3 类	0.860 4	0.790 1	0.815 7	0.675 9	0.712 9
D4 类	0.945 6	0.857 1	0.855 0	0.852 7	0.868 2

与产业类的文本较其他 5 个类别更易被区分,其次更易被分类的是动物类文本,而在 B、D1、D2、D3 类别上的表现并不突出,其值均在 0.85 ~ 0.9 之间,即科学与技术类、森林类、植物类、湿地类文章的部分样本具有相似性,这也与实际情况相符,林业的研究多集中于森林与湿地两大生态系统,并以林木花草等为研究主体,而在表 1 中也可以明显看到,“花卉”、“林木”、“研究”等词在类别中有交叉,与当前的实验结果相吻合。

在其他 4 种算法的结果中,GBDT 在 A 和 C 类的 F_1 高于 0.9,但在 D1 类低于 0.7;NB 算法在 A、D2、D4 类上的 F_1 达到 0.8 以上;ELM 与 BPNN 在各类别上的精度略有差距,两者整体的分类精度不高,说明在应对海量数据的分类场景时,两分类算法需进一步优化。

显然,XGBoost 算法在各个类别数据上的精准度、召回率以及 F_1 上有明显的优势,在优势类别中的文本分类精度极高,且该算法训练的模型更符合实际场景,从而验证了 XGBoost 在海量林业文本分类问题上的有效性。

图 4 是基于 Spark 并行环境下的各分类算法的准确率与执行效率的对比,其准确率与表 2 中 F_1 的分布一致:XGBoost 算法的分类准确率最高,为 0.923 4。从图 4b 可看出,BP 神经网络的训练时间最长,为 2 182 s,其次是 ELM 算法,这与神经网络算

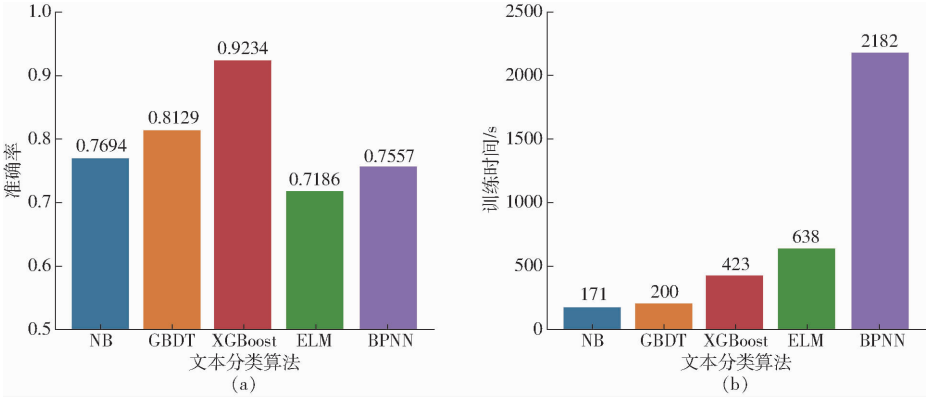

图 4 各算法的准确率与训练时间对比

Fig. 4 Comparison of accuracy value and training time of each parallel algorithm on 100% data

法的执行时间常高于传统机器学习算法的情形相符。因此，XGBoost 在保持极高准确率的情况下，时间执行效率方面远优于两种神经网络算法，略慢于传统机器学习算法。

3.4　不同训练集对分类结果的影响

依照林业文本分类流程，从语料的各个类别中分别随机取数量为 90、150、300、600、1 200、1 800、2 400、2 700、3 000 的样本构成信息量不同的实验数据集；其中，90% 用于模型训练，10% 用于测试，由交叉验证求取不同训练集下 XGBoost 最优分类结果如表 3 所示，并以同样流程求取其他算法的分类结果进行对比分析，对比结果如图 5 所示。

表 3　训练样本量对 XGBoost 算法 F_1 的影响

Tab. 3　Influence of number of training samples on XGBoost classification F_1 value

每类文本数(占比/%)	A 类	B 类	C 类	D1 类	D2 类	D3 类	D4 类
90(3)	0.947 4	0.909 1	0.916 7	0.666 7	0.666 7	0.833 3	0.947 4
150(5)	0.975 6	0.909 1	0.950 0	0.827 6	0.871 8	0.789 5	0.864 9
300(10)	1.000 0	0.833 3	0.971 4	0.925 4	0.921 1	0.857 1	0.920 0
600(20)	0.964 7	0.945 2	0.974 8	0.924 1	0.984 4	0.878 5	0.944 0
1 200(40)	0.974 6	0.919 5	0.984 3	0.864 0	0.933 3	0.984 3	0.961 1
1 800(60)	0.994 4	0.906 3	0.989 2	0.900 8	0.952 9	0.863 0	0.947 4
2 400(80)	0.994 1	0.914 2	0.975 7	0.874 2	0.923 5	0.884 5	0.926 9
2 700(90)	0.983 2	0.893 1	0.987 2	0.893 5	0.904 2	0.850 2	0.928 7
3 000(100)	0.998 4	0.896 3	0.982 9	0.890 0	0.907 2	0.860 4	0.945 6

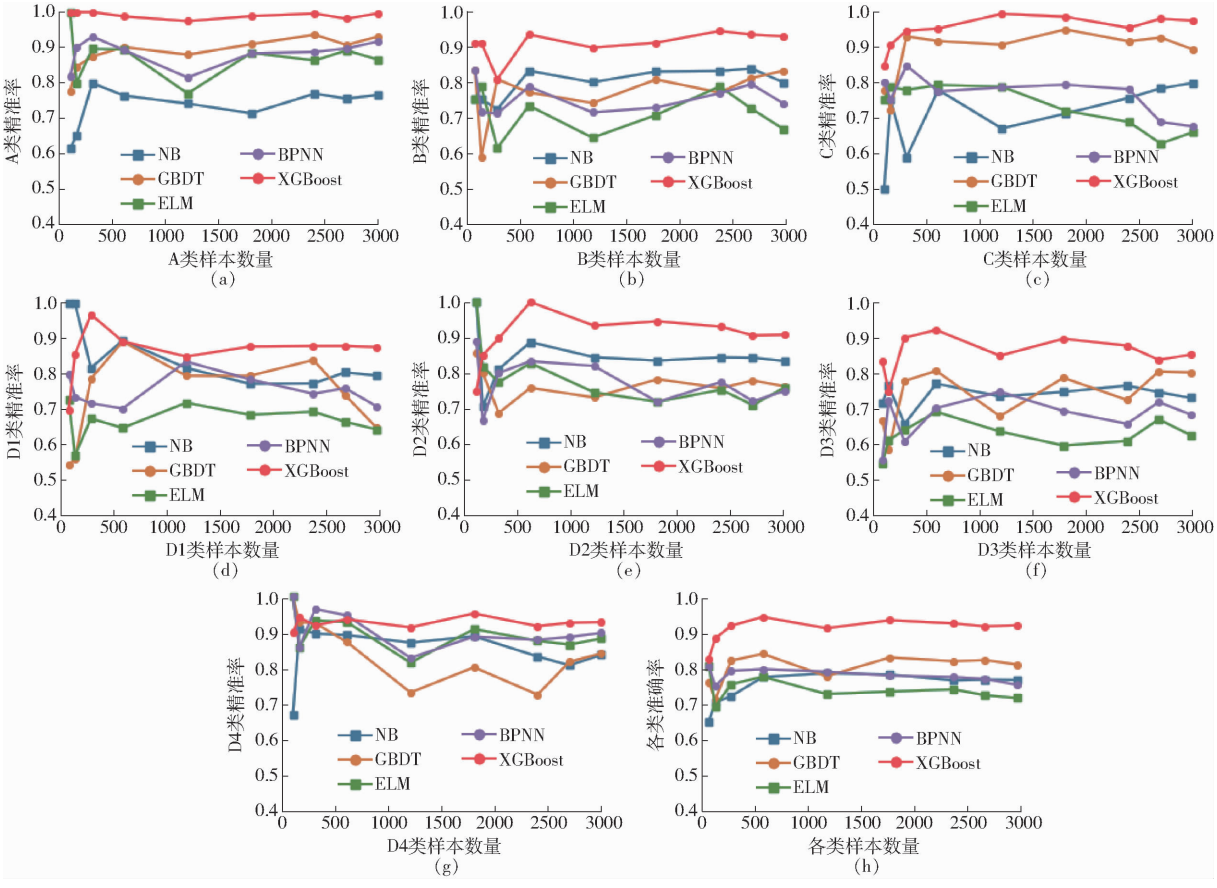

图 5　训练数量对各并行算法分类结果的影响

Fig. 5　Influence of number of training samples on results of each parallel algorithm

由表 3 与图 5 可以看出，样本数量对分类算法的精准率影响较大，在低于 600(20%)个训练样本下，各算法分类精度均不稳定，某些类别上的分类精准率会低于 0.7。随着训练样本数量的增加，XGBoost 精度稳定提升，而两类神经网络算法对新数据的支持能力较差，导致其精度不稳定。当样本文本数量增加到约 2 400(80%)个后，继续增加样本文本的数量，XGBoost 分类精准率提升缓慢直至趋于稳定；而 GBDT 和 NB 在多数类别上逐渐趋于稳定，少数类别伴随有轻微下减的趋势。由图 5h 可

见,随着样本的继续增加,XGBoost 算法的准确率扔保持缓慢上升的趋势。

最后,为验证 XGBoost 算法模型的实用性,选取中国林业网的 20 条最新新闻进行模型验证,验证集的准确率为 0.95,说明该分类器在实际场景下的林业文本分类应用性好,可直接用于互联网中的涉林文本的分类。

3.5 加速比

加速比通常用于衡量平台的计算节点数量对算法并行效率的影响。实验将计算节点数由单机模式逐渐增加工作节点到 4 个,将 2.1 万条、4.2 万条、8.4 万条训练集下的实验结果记录为表 4。

表 4 Spark 集群不同节点数对加速比的影响

Tab.4 Effect of number of distributed nodes on speedup ratio

工作节点数与加速比	数据量/万条					
	2.1		4.2		8.4	
	训练时间/s	测试时间/s	训练时间/s	测试时间/s	训练时间/s	测试时间/s
单机	902	88	3 982	177	11 091	238
1 个节点	930	96	4 255	187	11 996	259
4 个节点	423	31	1 147	47	2 906	72
加速比	2.13	2.83	3.47	3.77	3.82	3.31

由表 4 可以看出,在仅有一个工作节点的集群模式下,Spark 集群运行效率不及单机算法,原因在于 Spark 本身的资源调度需占用一部分资源和时间。在数据集仅有 2.1 万条时,加速比仅为 2.13,并不够明显;而增加 1 倍数据时,加速比提升至 3.47;增加至 4 倍时,加速比提升为 3.82。其中,当数据为 2.1 万条时,从单机至 4 个节点的运行时间分别为 902、930、423 s。可以看出,随着节点数的增加,实验所需要的训练时间呈下降趋势。

综上,该并行算法较单机版本效率提升明显,且数据量越大,该算法的并行效率越高。

4 结论

(1)针对现有林业分类研究中暴露出的分类标签设定不科学、实验训练产出的模型不具有实用性的问题,借鉴林业专家提出的林业主题信息种类,重新进行分类标签的设定;基于林业爬虫技术采集涉林文本,从林业需求出发,设计出分类粒度更细致的分类体系,使得分类模型可直接用于互联网中的海量涉林文本分类;将分类后的样本以统一的格式保存后,可逐渐积累林业语料,为后续层次更为细致的林业文本分类研究做铺垫。

(2)针对传统林业文本分类中执行效率低、精准度不高的问题,提出一种基于 Spark 计算环境的 XGBoost 并行化方法。各算法的对比结果表明,在包含不同比例的数据集上,该并行设计 XGBoost 算法较其他算法的优势表现在 3 个方面:① 各个类别的精度均高于其他算法,在优势类别上的 F_1 可达到 0.998 4。② 模型通过训练达到精度峰值所需的样本量较其他算法相对更少。③ 模型的精准率趋于稳定后,随着样本的增加,其精准率保持稳定缓慢增长,并未呈现出明显的下降趋势,适用于未来更多新语料加入模型进行训练的场景。此外,由加速比实验可以看出,该并行化算法较单机算法提升明显,且数据量越大,并行效率越高。综上,并行 XGBoost 算法可有效解决海量林业文本的高效、精准分类问题。

(3)本文的分类结果并未达到完全正确的水平,除受算法本身的限制外,类别之间的少量样本存在交叉现象也是原因之一。因此,本文建立的分类体系仍可以从细化分类粒度的层面加以改进。

参 考 文 献

[1] 刘广平,刘波,滕轶奕.“智慧林业”时代的信息资源开发与利用探讨[J].林业资源管理,2013(6):33 - 36.
LIU Guangping, LIU Bo, TENG Yiyan. Information resource development and utilization in the wisdom forestry times[J]. Forest Resources Management, 2013(6):33 - 36. (in Chinese)

[2] LI Zhao, LU Wei, SUN Zhanquan, et al. A parallel feature selection method study for text classification[J]. Neural Computing and Applications, 2017, 28(Supp.1): 513 - 524.

[3] 杨帅华,张清华. 粗糙集近似集的 KNN 文本分类算法研究[J]. 小型微型计算机系统, 2017, 38(10): 2192 - 2196.
YANG Shuaihua, ZHANG Qinghua. Research on K-nearest neighbor text classification algorithm of approximation set of rough set[J]. Journal of Chinese Computer Systems, 2017,38(10):2192 - 2196. (in Chinese)

[4] LIU Peng, ZHAO Huihan, TENG Jiayu, et al. Parallel naive Bayes algorithm for large-scale Chinese text classification based on Spark[J]. Journal of Central South University, 2019,26(1):1 - 12.

[5] YIN Chunyong, XI Jinwen, WANG Jin. The research of text classification technology based on improved maximum entropy model[C]//2015 First International Conference on Computational Intelligence Theory, Systems and Applications (CCITSA). IEEE, 2015.

[6] 魏芳芳,段青玲,肖晓琰,等.基于支持向量机的中文农业文本分类技术研究[J/OL].农业机械学报,2015,46(增刊):174 - 179.
WEI Fangfang, DUAN Qingling, XIAO Xiaoyan, et al. Classification technique of chinese agricultural text information based on SVM[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2015, 46(Supp.): 174 - 179. http://www.j-

csam. org/jcsam/ch/reader/view_abstract. aspx? file_no = 2015S029&flag = 1. DOI:10. 6041/j. issn. 1000-1298. 2015. S0. 029. (in Chinese)

[7] VATEEKUL P, KUBAT M. Fast induction of multiple decision trees in text categorization from large scale, imbalanced, and multi-label data[C]// IEEE International Conference on Data Mining Workshops, Miami, Florida, USA. IEEE, 2009.

[8] 赵明,杜会芳,董翠翠,等.基于 word2vec 和 LSTM 的饮食健康文本分类研究[J/OL].农业机械学报,2017,48(10):202 - 208.
ZHAO Ming, DU Huifang, DONG Cuicui, et al. Diet health text classification based on word2vec and LSTM [J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2017,48(10):202 - 208. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no = 20171025&flag = 1. DOI:10.6041/j.issn.1000-1298.2017.10.025. (in Chinese)

[9] 段青玲,魏芳芳,张磊,等.基于 Web 数据的农业网络信息自动采集与分类系统[J].农业工程学报,2016,32(12):172 - 178.
DUAN Qingling, WEI Fangfang, ZHANG Lei, et al. Automatic acquisition and classification system for agricultural network information based on Web data[J]. Transactions of the CSAE,2016,32(12):172 - 178. (in Chinese)

[10] WANG B, YIN J, HUA Q, et al. Parallelizing k-means-based clustering on spark[C]// 2016 International Conference on Advanced Cloud and Big Data (CBD). IEEE, 2016.

[11] 陈宇,王明月,许莉薇.基于 DE - ELM 的林业信息文本分类算法[J].计算机工程与设计,2015,36(9):2412 - 2415,2431.
CHEN Yu, WANG Mingyue, XU Liwei. Forestry information text classification algorithm based on DE - ELM[J]. Computer Engineering and Design, 2015,36(9):2412 - 2415,2431. (in Chinese)

[12] 陈宇,许莉薇.基于高斯混合模型的林业信息文本分类算法[J].中南林业科技大学学报,2014,34(8):114 - 119.
CHEN Yu, XU Liwei. Forestry information text classification algorithm based on GMM model [J]. Journal of Central South University of Forestry & Technology, 2014,34(8):114 - 119. (in Chinese)

[13] 陈宇,许莉薇.基于优化 LM 模糊神经网络的不均衡林业信息文本分类算法[J].中南林业科技大学学报,2015,35(4):27 - 32,59.
CHEN Yu,XU Liwei. Uneven forestry information text classification algorithm based on optimization LM fuzzy neural network [J]. Journal of Central South University of Forestry & Technology, 2015,35(4):27 - 32,59. (in Chinese)

[14] LIU Y, XU L, LI M. The parallelization of back propagation neural network in mapreduce and Spark[J]. International Journal of Parallel Programming, 2017,45(4): 760 - 779.

[15] 叶敏,汤世平,牛振东.一种基于多特征因子改进的中文文本分类算法[J].中文信息学报,2017,31(4):132 - 137,144.
YE Min,TANG Shiping, NIU Zhendong. An improved Chinese text classification algorithm based on multiple feature factors [J]. Journal of Chinese Information Processing,2017,31(4):132 - 137,144. (in Chinese)

[16] 叶雪梅,毛雪岷,夏锦春,等.文本分类 TF - IDF 算法的改进研究[J].计算机工程与应用,2019,55(2):104 - 109,161.
YE Xuemei, MAO Xuemin, XIA Jinchun, et al. Improved approach to TF - IDF algorithm in text classification[J]. Computer Engineering and Applications, 2019,55(2):104 - 109,161. (in Chinese)

[17] TORLAY L, PERRONE-BERTOLOTTI M, THOMAS E, et al. Machine learning-XGBoost analysis of language networks to classify patients with epilepsy[J]. Brain Informatics, 2017, 4(3):159 - 169.

[18] CHEN T, GUESTRIN C. XGBoost: a scalable tree boosting system[C]// ACM Sigkdd International Conference on Knowledge Discovery & Data Mining, 2016.

[19] 宋国琴,刘斌.基于 XGBoost 特征选择的幕课翘课指数建立及应用[J].电子科技大学学报,2018,47(6):921 - 926.
SONG Guoqin, LIU Bin. The establishment and application of drop outIndex of moocs based on XGBoost feature selection[J]. Journal of University of Electronic Science and Technology of China,2018,47(6):921 - 926. (in Chinese)

[20] 彭徵,王灵矫,郭华.基于随机森林的文本分类并行化[J].计算机科学,2018,45(12):148 - 152.
PENG Zheng,WANG Lingjiao, GUO Hua. Parallel text categorization of random forest[J]. Computer Science,2018,45(12):148 - 152. (in Chinese)

[21] 张丽莎,张贵,龙朝夕,等.林业专题动态信息的搜索与集成[J].中南林业科技大学学报,2013,33(5):47 - 51.
ZHANG Lisha, ZHANG Gui, LONG Zhaoxi, et al. Search and integration of thematic dynamic information on forestry [J]. Journal of Central South University of Forestry & Technology,2013,33(5):47 - 51. (in Chinese)

[22] 高军,张旭,刘燕,等.林业科技成果标识方法及其信息管理系统设计与实现[J].世界林业研究,2018,31(3):9 - 14.
GAO Jun, ZHANG Xu, LIU Yan, et al. Science and technology achievements identification and design & operation of its information management system[J]. World Forestry Research,2018,31(3):9 - 14. (in Chinese)

[23] 孙伟,曹姗姗,蒲智,等.林业资源信息云基础设施研究[J].西北林学院学报,2014,29(2):71 - 79.
SUN Wei, CAO Shanshan, PU Zhi, et al. Cloud infrastructure of forest resources information[J]. Journal of Northwest Forestry University,2014,29(2):71 - 79. (in Chinese)

[24] 袁津生,郭艳芬.林业主题爬虫的算法研究与设计[J].计算机工程与设计,2011,32(6):2003 - 2006.
YUAN Jinsheng, GUO Yanfen. Algorithm research and design of forestry focused web crawler[J]. Computer Engineering and Design,2011,32(6):2003 - 2006. (in Chinese)

[25] 王祥翔,方荟,陈崇成.基于朴素贝叶斯的文化旅游文本分类技术研究[J].福州大学学报(自然科学版),2018,46(5):644 - 649.
WANG Xiangxiang, FANG Hui, CHEN Chongcheng. Classification technique of cultural tourism text based on naive Bayes [J]. Journal of Fuzhou University(Natural Science Edition),2018,46(5):644 - 649. (in Chinese)