DOI: 10.11992/tis.201709011
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20180412.1032.006.html

大数据背景下高校招生策略预测

杨正理, 史文, 陈海霞, 王长鹏
（三江学院 机械与电气工程学院, 江苏 南京 210012）

摘　要: 在应届高中生生源不断下降、高等院校招生规模不断扩大、招生方式多元化不断发展、各院校之间招生竞争日趋激烈的条件下, 利用海量招生异构数据, 准确定位生源对象, 做好前期招生宣传是各高等院校需要考虑的重要问题。结合云计算技术, 利用并行化计算模型 MapReduce 和内存并行化计算框架 Spark 对高校招生历史数据进行分析, 提出采用并行化随机森林预测高校招生策略模型, 缩短了模型的预测时间、提高了模型的预测精度、增强了模型对大数据的处理能力。实验结果表明, 并行化随机森林算法在不同数据集上的多方面性能均优于常用的决策树预测方法。

关键词: 大数据; 机器学习; 深度学习; 学习算法; 高校招生; 策略预测; 随机森林; 云计算
中图分类号: TP311　　**文献标志码:** A　　**文章编号:** 1673−4785(2019)02−0323−07

中文引用格式: 杨正理, 史文, 陈海霞, 等. 大数据背景下高校招生策略预测[J]. 智能系统学报, 2019, 14(2): 323–329.
英文引用格式: YANG Zhengli, SHI Wen, CHEN Haixia, et al. The strategy of college enrollment predicted with big data[J]. CAAI transactions on intelligent systems, 2019, 14(2): 323–329.

The strategy of college enrollment predicted with big data

YANG Zhengli, SHI Wen, CHEN Haixia, WANG Changpeng
(School of mechanical and electrical engineering, SanJiang University, Nanjing 210012, China)

Abstract: Considering the decline in the enrollment of high school students and the expansion in the scale of enrollment of colleges and universities, methods of enrollment are developing continuously, and the competition among colleges and universities is becoming fierce. Under this background, an important issue that colleges and universities need to consider is to accurately locate the source of students by using the tremendous amount of heterogeneous enrollment data and accomplish the pre-enrollment propagation. Combined with the cloud computing technology, the parallel computing model MapReduce and the memory parallel computing framework Spark are used to analyze historical enrollment data. The paralleled random forest algorithm is proposed to predict the strategy of college enrollment. This model has a shorter prediction time, improved prediction accuracy, and improved big data processing ability. The experimental result shows that the performance of the paralleled random forest algorithm in different datasets is significantly superior to the widely used decision tree prediction method.

Keywords: big data; machine learning; deep learning; learning algorithm; college enrollment; strategy prediction; random forest; cloud computing

随着计算机通信网络技术、信息技术的发展, 普通高校招生方式多元化, 以及各院校招生竞争的日趋激烈, 制定精确合理的招生策略所需要参考的招生信息数据呈现爆炸性增长, 形成了招生信息大数据[1]。原有的招生信息数据处理方式已不能满足大数据的要求, 需要研究新的数据分析方法。

高校招生策略预测的常用方法有: 时间序列、灰色预测、多元统计等。这些方法具有简单实用、预测速度快的优点, 但只适用小样本、线性变化的数据集, 对大规模、非线性数据则无能为

收稿日期: 2017−09−11.　**网络出版日期:** 2018−04−12.
基金项目: 江苏省高校自然科学研究面上项目 (17KJB470011).
通信作者: 杨正理. E-mail: zhengli-yang@163.com.

力[2]。近年来，基于大数据技术，研究更有效的预测模型已成为学术界和产业界共同关注的热点[3]。文献[4]采用 Spark 平台和并行随机森林算法对短时电力负荷进行预测，改进了单机随机森林算法的各方面性能；文献[5]基于随机森林算法的并行化，对历史负荷数据及相关的温度、风速等一起进行分析，提高了负荷预测效率，并增强了算法对大数据的处理能力；文献[6]提出了一种基于弱相关化特征子空间选择的离散化随机森林并行分类算法，使决策树之间相关性降低，提高了随机森林的分类效果；文献[7]在小规模集群服务器上用消息传递技术对随机森林算法进行并行化，提高了模型的训练速度；文献[8]采用数据重构方法获取多维高校历史数据，利用非线性预测能力较强的支持向量机提出了一种数据挖掘高校招生预测模型；文献[9]以历年招生数据为基础，采用数据挖掘手段分析校园网络数据，构建了高校招生预测系统，为学校招生带来可视化的预测信息；文献[10]建立了高校招生数据挖掘系统，提出了有利于高校招生的策略预测方法。

经过众多研究学者的努力，国内对高校招生策略的预测方法取得了一定成果，但由于相适应的市场机制还没有形成，一些有效的预测模型，如并行化随机森林算法在高校招生领域还没有得到应用。本文借助 Hadoop 平台，利用并行化计算框架对招生数据进行挖掘和分析，提出了并行化的随机森林算法预测高校招生策略的方法。

1　大数据管理平台

1.1　Hadoop 技术

Hadoop 是云计算技术应用最广泛的平台之一，已经成为大数据管理与并行处理的主流技术。Hadoop 是一个开源的分布式软件框架，分布式文件系统 (hadoop distribution file system, HDFS) 和并行化计算模型 MapReduce 是其最核心内容[11]。HDFS 提供了文件分布式存储、大数据库管理等应用技术；而 MapReduce 则为大数据库提供了完善的并行分析计算框架。为了方便用户操作，Hadoop 还提供了一系列实用的组件供用户选择，如 Hive、Pig、Sqoop、Datanucleus 等[12]。

1.2　大数据管理平台框架结构

参照云计算技术体系结构[13]与数据分析处理工具，并结合高校招生数据分析的实际需要，搭建以数据存储、分析计算为主的高校招生数据管理平台，其基本构架如图 1 所示。平台自下往上分为：数据采集整合系统、数据存储系统、数据分析系统和数据应用系统。

该平台是 Hadoop 技术的具体应用：一方面，利用 Hadoop 的核心组件 HDFS、HBase、Hive 建立大数据存储系统；另一方面，利用 MapReduce 并行计算框架和 Spark 内存并行计算框架，构成数据计算分析系统，实现对高校招生数据的分析与计算。

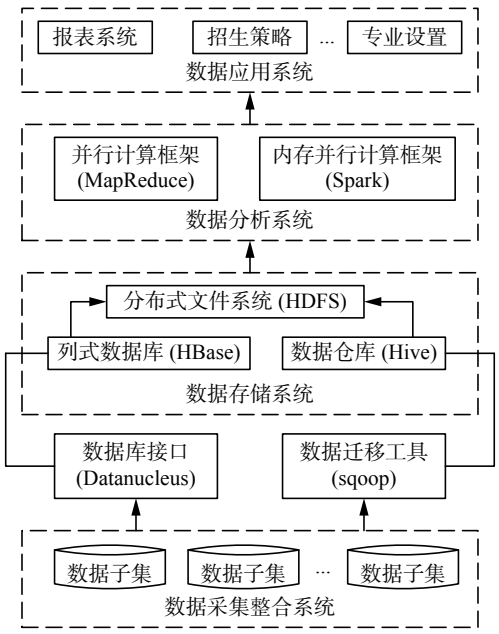

图 1　大数据管理平台框架图

Fig. 1　Architecture diagram of big data manage platform

1.3　数据采集整合系统

高校的招生人数、专业设置、生源人数、学生成绩等招生数据构成数据子集，这些数据子集来源不同，数据口径不一，模态千差万别，形成了海量异构数据。

数据整合过程就是将海量异构数据迁移至 Hadoop 集群，实现高效存储与管理。目前，数据整合过程还没有一个高效标准的方法，还需要利用第三方软件完成该操作，如 Sqoop、Datanucleus 等。Sqoop 能够将数据在 Hadoop 集群和关系型数据库之间进行相互转移[14]。在本管理平台中，利用 Sqoop 将各数据子集迁移到集群的数据仓库；Datanucleus 能够支持多种主流存储系统[15]，屏蔽各存储系统之间的差异，提供标准的数据接口 (JDO, JPA) 实现数据传送。在本管理平台中，各数据子集通过 Datanucleus 接口将数据导入到列数据库 HBase 中。

1.4　数据存储系统

数据仓库、列数据库中的数据均存储在 Hadoop 集群的 HDFS 中。采集到的原始数据经过抽取、清理、系统加工、整合等预处理后保存到数据

仓库, 预处理过程是为了保证数据仓库中的数据信息是一致的全局信息[16]。Hadoop 提供了一款管理数据仓库组件 Hive, 其作用是将结构化的数据文件映射成数据库, 并为用户提供简单的 SQL 查询功能[17]。HDFS 中的数据块 (Block) 采用冗余多备份机制存储, 能有效的处理单点故障。

1.5　数据挖掘分析系统

平台采用并行化计算模型 MapReduce 对数据进行挖掘分析, 利用基于内存的并行化计算模型 Spark 对对密集型数据完成迭代式计算。MapReduce 向用户提供了庞大但设计精良的并行计算软件框架, 在集群内能实现计算任务和数据的自动划分, 并能根据集群节点所能提供的资源自动完成任务的分配, 并有效监控任务的完成过程, 最后还能自动完成各集群节点计算结果的收集。MapReduce 将数据分布式存储、数据通信、容错处理等复杂的底层细节全交由系统处理, 大大减轻了用户软件开发负担[18]; Spark 是在 Hadoop 基础上进行改良的基于内存的集群计算系统。系统的中间数据全部存放在内存中, 对迭代等复杂的计算过程具有很高的效率[19]。

1.6　数据应用系统

根据云服务中应用即服务的概念, 数据应用系统就是向高校招生策略预测系统的应用者提供所需要的服务, 如以文件的形式提供各省市招生计划投放数据列表、指导本校专业设置建议、招生生源选择提示、招生宣传策略等可视化服务。数据应用系统还为用户提供与高校招生有关的、能够与其他系统进行数据交换的操作接口。

2　并行随机森林预测高校招生策略

2.1　随机森林算法原理

在大数据背景下, 常用的分类预测算法有极限学习、神经网络、遗传算法、支持向量机、决策树等。决策树在传统的分类预测算法基础上得到了广泛研究, 也取得了不错的应用效果[20], 但由于其自身原因, 仍然存在以下不足。

1) 在建树初始需要将所有的分类规则读入内存, 限制了决策树处理更多数据, 因此其处理大数据的能力有限。

2) 实际应用中, 当数据中有噪声或训练样本过少时, 会出现过度拟合现象。过度拟合的决策树对训练样本的分类效果表现良好, 但对新样本的分类效果则明显不佳。

3) 决策树在选择属性时不进行回归运算, 因此其结果仅能收敛于局部最优解, 造成决策树分类精度不高, 且泛化能力较差。

随机森林是一种集成了多棵分类回归树的综合分类预测算法。当输入训练样本时, 每一棵决策树都会产生一个分类结果, 通过对所有分类结果进行投票得到随机森林的最终分类结果。随机森林吸收了决策树的所有优点, 同时克服了决策树的缺点。又因为便于实现并行化, 提高了数据分析效率, 同时也提高了算法对大数据的处理能力。

由于高校招生策略的输出为实数, 只需要讨论随机森林的回归过程, 其实现步骤如下 (设集成的决策树棵数为R):

1) 从原始数据集S中采用 Bagging 方法有放回的抽取大小为N的训练子集$\mathrm{TS}_i(i=0,1,\cdots,R)$;

2) 对TS_i重复①~③步骤, 直到节点的样本数不超过预设的最小值$L_{\min}$, 得到一棵决策树T_i:

① 从M个属性样本集中随机抽取m个属性样本; 在回归模型中, m值取M的 1/3。

② 从m个属性样本中选择最佳的变量j和切分点s得到$\theta(j,s)$;

③ 将该节点$\theta(j,s)$切分成两个内部节点。

3) 所有决策树集合$\{T_i\}_1^R$构成随机森林。

决策树中内部节点进行分支的样本属性选择依据采用最小二乘偏差算法。采用“平方误差最小原则”来度量决策树的分支偏差, 节点t的拟合偏差公式为

$$\mathrm{Err}(t)=\frac{1}{n_t}\sum_{D_t}(y_i-k_t)^2 \tag{1}$$

式中: n_t为节点t中所包含的实例个数, k_t为每个内部节点中由实例目标值计算所得到的平均值。

节点t按属性值s进行分支的最小二乘偏差值计算公式为

$$\mathrm{Err}(s,t)=\frac{n_{tL}}{n_t}\mathrm{Err}(t_L)+\frac{n_{tR}}{n_t}\mathrm{Err}(t_R) \tag{2}$$

为了在训练过程中减少遍历属性值的计算, 对式 (2) 进行化简得到:

$$\mathrm{Err}(s,t)=\frac{S_L^2}{n_{tL}}+\frac{S_R^2}{n_{tR}} \tag{3}$$

式中: $S_L=\sum_{D_{tL}}y_i$, $S_R=\sum_{D_{tR}}y_i$, 划分的标准是使式 (3) 的计算值最大。

2.2　随机森林算法的并行化

随机森林集成了多个决策树, 这是随机森林算法能够实现并行化的物理条件。而袋装 (Bagging) 算法和随机子空间思想为随机森林算法的并行化提供了基本理论依据:

Bagging 算法是一种根据概率分布原理从数据集中有放回的抽样技术。Bagging 算法进行每轮抽样时，数据集中约有 36.8% 的样本不能被抽中，没有被抽中的数据样本不能参加算法训练，但可以用来检测训练模型的泛化能力。Bagging 算法使每个训练样本的内容不同，但所包含原始数据集的知识规模是相同的，从而使随机森林中的每个决策树的构建过程相互独立，可以并行完成训练过程。

随机子空间思想是指决策树在每个节点进行属性样本抽取时，随机的从属性样本中抽取若干个属性的方法。由于抽取过程随机，所以多个节点可以并行化地同步抽取，使各决策树可以独立生成。

Bagging 思想和随机子空间思想保证了随机森林能够并行运行，使其具有较高的预测精度、较快的数据分析效率和较强的数据处理能力。因此，本文提出了基于 MapReduce 的并行化随机森林算法 (MapReduce-paralleled random forests, MR-PRF) 进行高校招生策略预测方法。

3　并行随机森林算法实现

3.1　算法的预测流程

高校招生策略预测的原始数据量巨大，开启 3 个 MapReduce 作业类来完成数据处理过程。每个 MapReduce 类的输出作为下一个 MapReduce 类的输入，3 个 MapReduce 类分别完成生成数据字典、生成决策树和构建随机森林模型。

生成数据字典就是以文件的形式解析参于训练的样本数据，由第 1 个 MapReduce 作业类完成。在 Map 过程，首先读取一部分招生样本数据，然后提取样本数据的属性类型、属性值、以及模型的类型 (是回归还是分类)，得到 key/value 数据对传递给 Reduce 过程；在 Reduce 过程，将 Map 过程得到的 key/value 数据对按 key 值进行合并，并通过 Datanucleus 数据库接口写入到 HBase 中。所有的 key/value 数据对以文件形式进行记录，保存在集群的 HDFS 中，作为第 2 个 MapReduce 作业类的输入。

生成决策树由第 2 个 MapReduce 作业类完成。随机森林算法中集成的决策树是并行产生的，一个 Map 过程生成一个决策树。该 MapReduce 作业只有 Map 过程，没有 Reduce 过程。

生成随机森林由第 3 个 MapReduce 作业类完成。在回归预测模型中，该过程的主要功能就是将所有决策树的结果进行统计，求取平均值得到随机森林的最终结果。

采用并行化随机森林算法预测高校招生策略的具体流程如图 2 所示。该流程基于 Hadoop 集群强大的存储能力和数据处理能力，对招生数据进行挖掘和分析处理，有效的提高了算法的预测精度和数据处理能力。

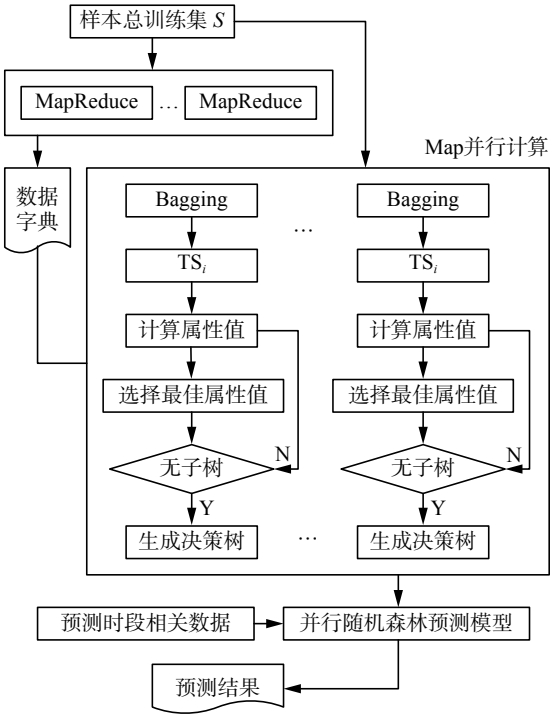

图 2　并行化随机森林招生策略预测流程图

Fig. 2　Flow chart of paralleled random forests for enrollment strategy

3.2　高校招生大数据实验平台

课题组在实验室采用 46 台计算机建立了一个高校招生策略预测实验平台。计算机集群采用典型的主/从结构，也称为 Master/Salve 结构。其中一台计算机作为 Master(管理节点)，负责集群内的资源管理和任务分配；其他计算机作为 Salve (数据节点)，负责保存各数据块，并完成与数据块相对应的任务。当 MapReduce 作业提交至 Master 节点时，Master 将数据文件进行分块，并记录与各数据块相对应的名字空间与元数据。然后将各数据块冗余的保存在各数据节点并分配相应的作业任务，并负责监控 MapReduce 作业的执行过程。实验平台的拓扑结构如图 3 所示。

图 3 中，大数据库以关系型数据库方式保存，应用 Sqoop 软件将本地文件或数据库表与 HDFS 文件进行相互迁移。Sqoop 软件是基于 MapReduce 实现的，用户无需过多关注 MapReduce 的实现和优化过程。实验中，将约 20 万条测试数据整合到 HBase 列式数据库中，大约需要 2 min 时间。

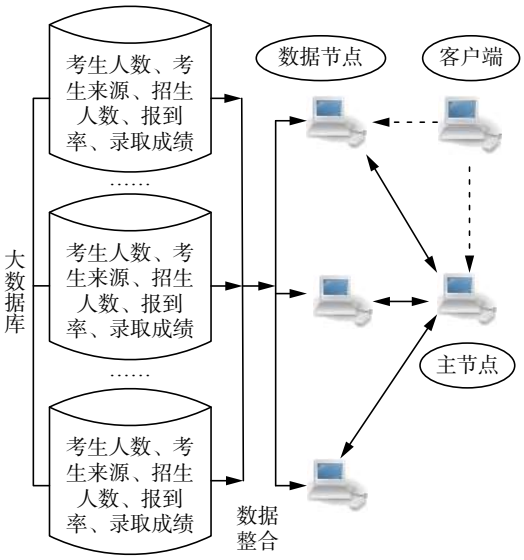

图 3　实验平台拓扑结构

Fig. 3　Topology map of experimental platform

3.3　实验数据、属性值、实验评价指标选取

实验数据来自某高校近 3 年的招生数据, 包括: 该年各省考生人数、考生来源 (毕业中学、中学所在地)、各专业在各省的招生人数、报到率、录取志愿排名、男女比例、学生当年录取成绩 (总分、选测成绩)、录取成绩在本省排名等。已有的数据远没有达到大数据库的规模, 但采用这些数据足以验证算法的正确性。后期通过人为的补充数据操作, 使实验数据达到大数据的规模, 然后验证算法的数据处理能力。根据大量文献[21-24]的研究成果, 将预测当年的招生数据进行归一化处理, 形成预测高校招生策略的样本属性。

算法的预测精度采用平均绝对百分比误差 (mean absolute percentage error, MAPE) 来评价, MAPE 的计算方法为

$$\mathrm{MAPE}=\left[\sum_{t=1}^{n}(|Y_t-y_t|/y_t)\right]/n\times 100\% \tag{4}$$

式中: Y_t为算法的预测值; y_t为真实值; n为预测结果的个数; MAPE 值越小时, 说明算法的预测精度越高。

算法的加速比 (speedup) 是指单位任务在单处理器系统下执行完成所消耗时间与该任务在并行处理器系统下执行完成所消耗时间的比值, 其作用是用来评价并行系统或程序并行化的性能和效果, speedup 的计算公式为

$$S_p=t/T \tag{5}$$

式中: t为单台计算机的运行时间, T为集群模型的运行时间

4　实验结果分析

实验 1　在相同的数据集下, 比较 MR-PRF 算法、决策树算法、单机随机森林算法的性能。原始数据集取 2014—2016 年某大学的历史招生数据 (文件大小为 104 MB, 共 1.2×10^6 条数据), 分别采用 MR-PRF 算法 (集成决策树数量$R=240$)、决策树算法、单机随机森林算法对 2017 年的招生策略进行预测, 各类实验均进行多次, 并取实验结果的平均值作为最终结果, 实验结果如表 1 所示。

表 1　各类算法的预测性能比较

Table 1　Prediction performance of all kinds of algorithms

预测算法	MAPE 值/%	运行时间/s
MR-PRF 算法	1.39	7.5
决策树算法	2.15	241
单机随机森林	4.19	369

由表 1 可见, MR-PRF 算法的预测性能最好, 且执行效率最高。这是因为 MR-PRF 算法吸取了决策树的优点而克服了其缺点, 在预测精度上才有更好的表现。而且由于 MR-PRF 算法的并行化, 使其执行效率得到较大提高。

实验 2　在同样的数据集下, MR-PRF 算法集成决策树的数量R与算法的性能表现之间的关系。采用实验 1 数据集, MR-PRF 集成决策树数量R取不同值时, 得到的实验结果如表 2 所示。

表 2　MR-PRF 算法的预测精度受决策树数量的影响

Table 2　The prediction accuracy of the MR-PRF algorithm is affected by the number of decision trees

决策树数量	MAPE 值/%	运行时间/s
120	2.16	2.8
180	1.93	4.7
240	1.61	7.3
300	1.58	9.6
360	1.59	12.9

由表 2 可见, MR-PRF 算法的集成决策树数量R取值过小时, 算法精度较低, 这是因为不能充分体现 MR-PRF 的并行优势; 当 MR-PRF 算法的集成决策树数量R取值过大时, 算法的复杂程度加大, 预测时间加长; 当 MR-PRF 算法的集成决策树数量R取值达到一定程度时, 算法的精度变化不大。这说明在实际应用时, R取值应合理。

实验 3　MR-PRF 算法的集成决策树数量 R 取值一定时 ($R=240$), 其预测性能和数据集大小的关系。人为补充数据集至不同大小, 对每组数据集分别进行多次实验, 取多次实验的平均值作为最终结果, 实验数据如表 3 所示。

由表 3 可见，原始数据集的大小对 MR-PRF 算法的预测性能影响不大，没有明显的规律可寻。但随着原始数据集的增加，运行时间加大，这是符合算法规律的。该实验结果表明 MR-PRF 算法是适合处理大数据集的。

表 3　MR-PRF 算法的预测性能受数据集大小的影响

Table 3　The prediction property of the MR-PRF algorithm is affected by the data set size

文件大小 (MB)/元组数	MAPE 值/%	运行时间/s
340/4.3×10^6	1.74	36.3
680/8.6×10^6	1.72	88.1
1 020/12.9×10^6	1.67	129.6
1 360/17.2×10^6	1.71	177.8
1 700/21.5×10^6	1.65	241.7

实验 4　通过计算加速比值来评价 MR-PRF 算法的并行性能。人为补充数据集至 3.6、13.6、136 GB，分别由 1、5、15、25、35 台计算机构成集群，选择 MR-PRF 算法集成决策树数量 $R = 240$ 进行预测实验，结果如图 4 所示。由图 4 可见，在相同规模集群下，数据集越大，加速比越大，并行性能越好；在相同的原始数据集下，加速比随集群的增加而增加，并行性能也越好。

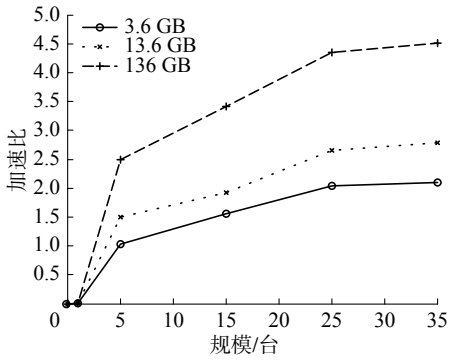

图 4　MR-PRF 算法的加速比

Fig. 4　Speedup of MR-PRF algorithm

5　结束语

在国内外大数据研究基础上，针对高校招生数据集的特点，提出了一种基于 Hadoop 的分布式并行随机森林算法模型，并利用该模型处理高校招生大数据，实现对未来招生策略进行预测。经多次不同类型的实验进行验证，并与使用广泛的决策树预测算法进行比较，证明并行随机森林算法模型具有更快的数据分析速度，更高的预测性能以及更好的大数据处理能力。

受实验条件限制，原始招生数据集在数量上远没有达到大数据的规模，但通过人为的数据补充操作，提高了实验的真实性。因此，本文的结论仍然具有较强的可参考性。

参考文献：

[1] TOLLE K M, TANSLEY D S W, HEY A J G. The fourth paradigm: data-intensive scientific discovery[J]. Proceedings of the IEEE, 2011, 99(8): 1334–1337.

[2] MAYER-SCHONBERGER V, CUKIER K. Big data: a revolution that will transform how we live, work and think[M]. Boston: Hodder Press, 2013.

[3] RUSITSCHKA S, EGER K, GERDES C. Smart grid data cloud: a model for utilizing cloud computing in the smart grid domain[C]//Proceedings of the First IEEE International Conference on Smart Grid Communications. Gaithersburg, MD, USA, 2010: 483–488.

[4] 刘琪琛, 雷景生, 郝珈玮, 等. 基于 Spark 平台和并行随机森林回归算法的短期电力负荷预测[J]. 电力建设, 2017, 38(10): 84–92.
LIU Qichen, LEI Jingsheng, HAO Jiawei, et al. Short-Term power load forecasting based on spark platform and parallel random forest regression algorithm model[J]. Electric power construction, 2017, 38(10): 84–92.

[5] 王德文, 孙志伟. 电力用户侧大数据分析与并行负荷预测[J]. 中国电机工程学报, 2015, 35(3): 527–537.
WANG Dewen, SUN Zhiwei. Big data analysis and parallel load forecasting of electric power user side[J]. Proceedings of the CSEE, 2015, 35(3): 527–537.

[6] 陈旻骋, 袁景凌, 王啸岩, 等. 基于弱相关化特征子空间选择的离散化随机森林并行分类算法[J]. 计算机科学, 2016, 43(6): 55–58, 90.
CHEN Mincheng, YUAN Jingling, WANG Xiaoyan, et al. Parallelization of random forest algorithm based on discretization and selection of weak-correlation feature subspaces[J]. Computer science, 2016, 43(6): 55–58, 90.

[7] 程光, 王贵锦, 何礼, 等. 人体姿势估计中随机森林训练算法的并行化[J]. 计算机应用研究, 2014, 31(5): 1558–1561, 1576.
CHENG Guang, WANG Guijin, HE Li, et al. Parallelization for randomized forests used in human pose estimation [J]. Application research of computers, 2014, 31(5): 1558–1561, 1576.

[8] 孙晓莹, 郭飞燕. 数据挖掘在高校招生预测中的应用研究[J]. 计算机仿真, 2012, 29(4): 387–391.
SUN Xiaoying, GUO Feiyan. Research on data mining for college enrolment prediction[J]. Computer simulation, 2012, 29(4): 387–391.

[9] 韩娜, 廖晨, 许杰维, 等. 基于大数据的高校招生预测系统的设计与实现[J]. 信息技术, 2016(12): 80–83.
HAN Na, LIAO Chen, XU Jiewei, et al. Design and implementation of college enrollment forecasting system based on big data[J]. Information technology, 2016(12): 80–83.

[10] 朱丽丽. 数据挖掘在高校招生中的应用研究[J]. 计算机与现代化, 2012(8): 190–194.
ZHU Lili. Research on application of data mining techno-

logy in enrollment of vocational colleges[J]. Computer and modernization, 2012(8): 190–194.

[11] 马世龙, 乌尼日其其格, 李小平. 大数据与深度学习综述[J]. 智能系统学报, 2016, 11(6): 728–742.
MA Shilong, WUNIRI Qiqige, LI Xiaoping. Deep learning with big data: State of the art and development[J]. CAAI transactions on intelligent systems, 2016, 11(6): 728–742.

[12] 龚冬颖, 黄敏, 张洪博, 等. RGBD 人体行为识别中的自适应特征选择方法[J]. 智能系统学报, 2017, 12(1): 1–7.
GONG Dongying, HUANG Min, ZHANG Hongbo, et al. Adaptive feature selection method for action recognition of human body in RGBD data[J]. CAAI transactions on intelligent systems, 2017, 12(1): 1–7.

[13] 张钢, 谢晓珊, 黄英, 等. 面向大数据流的半监督在线多核学习算法[J]. 智能系统学报, 2014, 9(3): 355–363.
ZHANG Gang, XIE Xiaoshan, HUANG Ying, et al. An online multi-kernel learning algorithm for big data[J]. CAAI transactions on intelligent systems, 2014, 9(3): 355–363.

[14] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. Computer science, 2015.

[15] 孟祥萍, 周来. 基于 hadoop 云平台的智能电网 HDFS 资源存储技术研究[J]. 电测与仪表, 2014, 51(19): 23–30.
MENG Xiangping, ZHOU Lai. Research on resource storage technologies of HDFS for smart grid based on hadoop cloud platform[J]. Electrical measurement & instrumentation, 2014, 51(19): 23–30.

[16] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484–489.

[17] 冯兴杰, 吴稀钰, 赵杰, 等. QAR 数据仓库在 Hive 中的构建[J]. 计算机工程与应用, 2017, 53(11): 90–94.
FENG Xingjie, WU Xiyu, ZHAO Jie, et al. Data warehouse of QAR based on hive[J]. Computer engineering and applications, 2017, 53(11): 90–94.

[18] 马学森, 王晓洁, 韩江洪, 等. MapReduce 框架下的 Skyline 结果优化算法[J]. 传感器与微系统, 2017, 36(2): 146–149.
MA Xuesen, WANG Xiaojie, HAN Jianghong, et al. Skyline result optimization algorithm based on MapReduce framework[J]. Transducer and microsystem technologies, 2017, 36(2): 146–149.

[19] 李帅, 吴斌, 杜修明, 等. 基于 Spark 的 BIRCH 算法并行化的设计与实现[J]. 计算机工程与科学, 2017, 39(1): 35–41.
LI Shuai, WU Bin, DU Xiuming, et al. Design and implementation of BIRCH algorithm parallelization based on Spark[J]. Computer engineering & science, 2017, 39(1): 35–41.

[20] 黄春华, 陈忠伟, 李石君. 贝叶斯决策树方法在招生数据挖掘中的应用[J]. 计算机技术与发展, 2016, 26(4): 114–118.
HUANG Chunhua, CHEN Zhongwei, LI Shijun. Application of Bayesian decision tree method in admission data mining[J]. Computer technology and development, 2016, 26(4): 114–118.

[21] 李战怀, 王国仁, 周傲英. 从数据库视角解读大数据的研究进展与趋势[J]. 计算机工程与科学, 2013, 35(10): 1–11.
LI Zhanhuai, WANG Guoren, ZHOU Aoying. Research progress and trends of big data from a database perspective[J]. Computer engineering & science, 2013, 35(10): 1–11.

[22] 吴倩红, 高军, 侯广松, 等. 实现影响因素多源异构融合的短期负荷预测支持向量机算法[J]. 电力系统自动化, 2016, 40(15): 67–72, 92.
WU Qianhong, GAO Jun, HOU Guangsong, et al. Short-term load forecasting support vector machine algorithm based on multi-source heterogeneous fusion of load factors[J]. Automation of electric power systems, 2016, 40(15): 67–72, 92.

[23] 陶永才, 丁雷道, 石磊, 等. MapReduce 在线抽样分区负载均衡研究[J]. 小型微型计算机系统, 2017, 38(2): 238–242.
TAO Yongcai, DING Leidao, SHI Lei, et al. Research on MapReduce on-line load balancing based on sample partition[J]. Journal of Chinese computer systems, 2017, 38(2): 238–242.

[24] 黄有福. 数据挖掘技术在招生数据平台的应用研究[J]. 电脑知识与技术, 2015, 11(31): 3–4.
HUANG Youfu. Application of data mining technology in the enrollment data platform[J]. Computer knowledge and technology, 2015, 11(31): 3–4.

作者简介:

杨正理, 男, 1971 年生, 副教授, 主要研究方向为复杂系统与计算智能、软件工程。参与 2 个省部级项目。发表学术论文 40 余篇。

史文, 女, 1983 年生, 讲师, 主要研究方向为云计算与大数据、计算机软件形式化方法。参与 2 个省部级项目。发表 10 余篇学术论文。

陈海霞, 女, 1978 年生, 副教授, 主要研究方向为海量信息处理的计算模型、自动推理。参与 3 个省部级项目。发表 20 余篇学术论文。