DOI:10.13718/j.cnki.xsxb.2021.01.001

基于增强可伸缩随机森林的高维大数据预测分析系统①

李发陵,　彭　娟

重庆工程学院 软件学院,重庆 400056

摘要:针对大数据由于数据复杂性、异构性、安全性、可伸缩性和大规模数据量而难以预测分析的问题,提出了基于增强可伸缩随机森林(Enhancing Scalable Random Forest, ESRF)的高维大数据预测分析系统. 该系统通过在训练数据集上执行超参数优化来提升可伸缩随机森林(Scalable Random Forest, SFR),然后对预处理数据应用主成分分析(Principal Component Analysis, PCA)和信息增益(Information Gain, IG),对不影响模型的特征进行缩减以减少模型开发阶段的处理时间开销. 实验结果表明,本文系统可以提供出色的预测能力,而且可以在整个实验数据集中以最少的处理时间提供有效的性能.

关　键　词:高维大数据;增强可伸缩随机森林;降维;预测分析;超参数优化

中图分类号:TP393　　**文献标志码**:A　　**文章编号**:1000-5471(2021)01-0001-06

随着信息技术的高速发展,每天产生的数据正以前所未有的速度增长和累积,如何更加高效地处理大数据成为当前研究的热点[1-2]. 大数据是指包括异构格式的海量数据增长,它的庞大性导致传统的数据处理软件在大数据应用的情况下不再有效[3-4]. 大数据分析检查来自不同来源的海量而多样的数据集以发现包括隐藏模式、未知相关性和其他影响业务决策的信息[5-7],它需要可伸缩的体系结构来存储和处理任何大小的数据,而不消耗大量的资源.

预测性大数据分析(Predictive big data analytics, PBA)是一种数据驱动技术,它可以分析大规模数据以发现模式和机会并预测结果[8-9],PBA 使用机器学习算法分析现在和以前的数据来预测未来的事件. Hadoop 预测用于在分布式群集数据管理系统中存储和处理大型异构数据集,Hadoop 分布式文件系统(Hadoop Distributed File System, HDFS)是提供对应用程序数据高性能访问的主要数据存储系统[10],Hadoop MapReduce 为预测分析提供基于纱线的海量数据并行处理,但不能有效地处理迭代和交互式分析任务. 为了高效地支持迭代任务,Apache Spark 引入弹性分布式数据集(Resilient Distributed Datasets, RDD),用于跨一组节点缓存 Spark 作业的中间数据,Spark 由于其众多的优点而成为高级分析系统中的一块基石[11].

在降低数据集维度和维护数据完整性之间保持平衡是大数据预测分析需要解决的关键问题. 文献[12]提出基于 Apache Spark 平台的大数据并行随机森林算法,该算法基于数据并行和任务并行优化相结合的混合方法进行优化,以解决高维数据问题,但是在进行降维时,PCA 和 IG 方法会影响 PBA 系统的预测性

① 收稿日期:2020-01-06
基金项目:国家自然科学基金项目(61572089);重庆工程学院科研基金资助项目(2019xzky02);重庆市教育科学项目(2017-GX-038).
作者简介:李发陵,硕士,副教授,主要从事软件工程、大数据存储与分析研究.

能. 文献[13]提出一种改进的随机子集特征选择算法来处理高维大数据，该算法使用标准的 k 最近邻分类器进行分类，基于随机森林算法通过选择有用的新特征来降低数据集的维数，改善了现有算法的降维性能，并增加了稳定性. 文献[14]提出一种使用最优功能基于可伸缩随机森林(Scalable Random Forest，SRF)的大数据分类方法，该方法使用改进的蜻蜓算法从数据库中选择最优属性以获得更好的分类效果，利用 SRF 分类器结合最优特征对电子健康数据进行分类.

为了进一步提高分类准确性和稳定性，本文提出基于增强可伸缩随机森林(Enhancing Scalable Random Forest，ESRF)的高维大数据预测分析系统，用于处理高维的海量数据. 本文将 SRF 中的超参数优化与降维相结合，显著提高系统的预测性能. 该系统使用贪婪方法寻找 SRF 的最优超参数组合，从高维大数据中预测趋势和行为模式，使用 PCA 和 IG 技术来减少数据集的不相关特征变量，以避免模型开发阶段的处理时间开销. 实验结果表明，本文的 PBA 系统可以提供出色的预测能力，而且在整个实验数据集中以最少的处理时间提供有效的性能.

1 设计的预测大数据分析系统

本文系统通过执行超参数优化来增强 SRF，通过对数据降维来改善预测分析的性能. 为实现预测分析平台，将 HDFS 用于分布式存储，将 Apache Spark 用于并行计算.

1.1 本文系统的体系结构

本文系统由数据存储、数据处理和数据分析组成.

数据存储检索大量的数值数据，并连接到支持快速处理的计算服务器节点. HDFS 用于提供具有容错方式的可扩展存储，当 HDFS 接收大量数据时，它将信息分解为单独的块，并将它们分发到集群中的不同计算节点，旨在为大数据提供可扩展且经济高效的存储. 此外，它是专为高度容错率而开发的，支持复制数据和将副本分发到不同的服务器上，因此可以在集群内的其他位置找到节点上的粉碎数据，确保在恢复数据的同时可以继续处理.

数据处理是本文系统获得先进计算基础设施最重要的部分之一，可以及时有效地挖掘和分析大规模数据. Apache Spark 处理引擎用于为大规模数据处理提供快速可靠的引擎，具有高级的有向无环图(Directed Acyclic Graph，DAG)执行引擎，支持无环数据流和内存计算，可以连接到几种类型的 Yarn 集群管理器，这些管理器在整个应用程序中分配资源. Spark Context 是将任务发送到执行程序以运行的主要组件，Spark 提供将数据缓存在内存中的功能，以便直接从内存中对相同的数据执行计算和迭代，因此节省了大量的磁盘 I/O 操作时间.

数据分析是本文系统获得高度预测准确性的重要组成部分，包括数据预处理和系统预测模型开发两个阶段. 现实世界中的大数据通常可能包含不一致和冗余的数据，执行数据清理步骤通过应用平滑技术来减少噪声，消除异常值. 使用标准归一化技术执行数据转换和还原步骤，有助于提供更容易理解的预测模式，执行相关性分析以检测对预测任务没有贡献的冗余属性. 在预测模型开发阶段，SRF 为本文系统提供准确的决策，是著名的 PBA 系统预测因子.

SRF 是树预测器与随机特征子空间和袋装法相结合的集成，它从学习样本中创建引导集，并学习相应的随机树. 在预测期间，森林中的单元树为预测投单独的票，然后使用聚合方法将这些票合并，以计算最终的整体预测. 由于 SRF 由多路复用和相互关联的计算组成，在计算大数据时会引起麻烦. 当学习技术没有优化和安排时，每个学习过程的中间和输出结果可能成为所有分布式设置的瓶颈. 本文系统使用来自 Spark MLlib 的随机森林(Random Forest，RF)模型构造，该算法在给定超参数训练 RF 模型的阶段将默认参数与标准参数一起使用，但对于高速、复杂和可变的数据，默认值无法获得最佳的预测精度. 此外，一个数据库不能够满足所有数据性质，在 SRF 中进行超参数优化以提高 PBA 系统的性能.

1.2　超参数优化

SRF 需要超参数的最优值来高效地处理大数据. 尽管默认值可用于某些预测，但它对所有数据集都是次优的，为了进行这一假设，通过执行超参数优化来增强 SRF. 超级参数是 SRF 的参数设置，可以在训练前调整以优化性能. SRF 降低了学习模型的复杂性，并将过拟合风险降至最低.

1.3　降维

数据集维度指的是数据集中呈现的特征数量，为了在分布式环境下进行有效的分析，降低数据维度已经成为一项重要的任务. 为了验证降维技术的有效性，本文评估了两种流行的特征降维技术：主成分分析(Principal Component Analysis，PCA)和信息增益(Information Gain，IG).

1.3.1　使用 PCA 进行降维

PCA 是一种通过将大数据投影到一个捕获大部分变化的子空间中来降低大数据高维性的特殊技术，主要目标是通过用一组称为主成分的新变量来找到 k 维. 数据中的大部分变化很大程度上是由第一个新的主成分捕获的，这意味着第一个分量是最重要的主轴，它具有最大的数据点方差. 当数据的变量彼此共线时，PCA 可以提供最佳结果，重要的是设置正确的主成分(Principal Component，PC)的最佳 k 个数.

为了找到 PCA，首先计算矩阵 X 的特征值分解. 设数据集为$(n\times d)$ 矩阵 X，S_i 是给定$(n\times d)$ 矩阵 X 的数据子集，将目标维数设置为 k，并计算$(n\times k)$ 矩阵 A，其中 n 是实例数，d 是原始维数，X 的列是矩阵 X 的主分量. 数据点的协方差矩阵 X_c 是由 M_c 的平均中心矩阵及其转置相乘得到，通过从 X 的每一行中减去矩阵 X 所有列均值的向量来计算M_c，然后计算特征向量和相应的特征值. 特征向量根据它们的特征值按降序排列，最高主成分是具有最高特征值的特征向量，选择前 k 个特征向量作为新 k 维的形式. 因此，PCA 将矩阵 X 的原始 n 维空间转换为新的 k 维.

1.3.2　使用 IG 进行降维

特征选择技术 IG 通过发现原始数据的特征重要性来降低实验数据集的维数，该方法的基本思想是通过计算每个特征变量的增益比值来对属性进行排序，最上面的特征变量是主变量，然后从剩余的变量中选择其他变量. 使用 IG 可以很容易地选择最大值，为了避免过度拟合问题，采用增益比值取最大值的特征变量.

使用 IG 对特征进行降维应标准化训练子集 S_i 和计算目标特征变量的熵. 对训练数据集 S 中的每个变量 y_{ij}，计算每个变量的熵、分裂节点信息、信息增益值、增益比数值和特征重要性. 特征向量根据它们的特征值按降序排列，最高主成分是具有最高特征值的特征向量，选择前 k 个特征向量作为新 k 维的形式. 数据集子集的信息熵定义为

$$Entropy(S_i) = -\sum_{c=1}^{d} q_c \times \log q_c \tag{1}$$

式(1) 中：S_i 是数据集 S 的子集，d 是 S_i 中不同类的数量，q_c 是 S_i 中属于类 c 的概率案例的数量. 子集 S_i 的每个变量 y_{ij} 的信息熵定义为

$$Entropy(y_{ij}) = \sum_{v\in V(y_{ij})}^{d} \frac{|S_{(v,i)}|}{|S_i|} Entropy(v(y_{ij})) \tag{2}$$

式(2) 中：y_{ij} 是子集数据集 S_i 的第 j 个输入特征变量，$v(y_{ij})$ 是每个输入变量的所有可能值的集合，$S_{(v,i)}$ 是数据集 S_i 的样本子集. 每个输入变量的分裂信息定义为

$$I(y_{ij}) = \sum_{c=1}^{m} -q_{c,j} \times \log(q_{c,j}) \tag{3}$$

式(3) 中：m 是 y_{ij} 不同类别的数目，$q_{c,j}$ 是属于类 c 的 y_{ij} 概率案例的数目. 每个特征变量的信息增益值定义为

$$G(y_{ij}) = Entropy(S_i) - Entropy(y_{ij}) \tag{4}$$

通过减少不影响预测能力的特征来最小化 SRF 的训练时间，IG 降维用于减少训练样本，这会降低预测

器的性能. 使用特征的熵来计算变量的增益比，并测量特征重要性. 因此，在SRF中考虑特征重要性对于高维大数据非常有效.

1.4 ESRF模型构建

将输入数据集分为用于建立预测模型的训练集和用于验证生成的模型测试集，在训练数据集上进行SRF的超参数优化以进行有效的模型开发，然后对预处理后的数据进行降维，最后基于最优参数建立ESRF预测模型.

ESRF建模过程输入具有大小为 N 和 M 要素的数据集 S、随机森林中树的数量 P_{nt} 和每棵树的最大深度 P_{md}. 它首先预处理来自 S 的数据，创建空字符串数组 F_i，将预处理数据分为训练数据集 S_{tr} 和测试数据集 S_{test}，从 S_{tr} 中减少不相关的特征和创建引导样本$\{S_{tr(1)}, S_{tr(2)}, \cdots S_{tr(nt)}\}$. 然后，定义一组超参数集$\{nt = 1, 2, \cdots nt_{\max}; md = 1, 2, \cdots md_{\max}\}$，结合每一组超参数$(nt, md)$生成RF模型. 在 S_{test} 上拟合模型，基于 M 特征预测结果，预测回归评估值，将误差值设为 $F_i\{0, 1, 2, \cdots, |nt_{\max}| \times |md_{\max}|\}$，按升序对 F_i 的误差值进行排序得到模型集$\{M_1, M_2, \cdots, M_{nt}\}$.

评估预测模型的性能，并为相应的数据集挑选精度最高的最佳预测模型，在执行模型选择后，使用测试数据集对所选模型进行验证，以正确预测大规模数据.

2 实验与结果分析

所有实验均在分布式大数据分析平台上进行，该平台由1个主节点和3个工作节点实现，单个节点在具有Intel(R)Core(TM)i7-6500U处理器和8.00 GB内存的Linux Ubuntu-16.04系统中执行，软件组件是Apache Hadoop发行版2.7.1、Spark 2.2.0和Scala 2.11.8. 为了评估本文PBA系统的性能，在实验中使用来自并行工作负载档案(Parallel Workload Archives，PWA)和加州大学尔湾分校(University of California Irvine，UCI)提出的用于机器学习的数据库的4个实际数据集.

为了衡量系统的预测准确性，使用平均绝对误差(*MAE*)和均方根误差(*RMSE*)作为评估指标，其定义如式(5)和式(6)所示.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - y_i'| \tag{5}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_i')^2}{n}} \tag{6}$$

其中，y_i 是数据集 S_i 输入特征变量的预测值，y_i' 是数据集 S_i 输入特征变量的实际值，n 为数 S_i 输入特征变量的数量.

在回归问题中测量模型评估的异常值是有用的，*MAE* 对异常值的存在具有鲁棒性，而 *RMSE* 则没有. *MAE* 和 *RMSE* 的范围从0～+∞. 误差值越低模型越好.

表1给出了不同数据集下SRF中默认参数和优化超参数的 *MAE* 比较结果. 超参数优化与这些数据性质中的 *MAE* 值密切相关. 密切关注特定参数，可以看出执行超参数优化可以提供比默认参数设置更好的精度.

表1 各数据集默认参数和超参数优化的 *MAE* 比较

数据集	*MAE*(默认参数)	*MAE*(优化参数)
DAS	1.418 2	0.519 1
HPC	1.547 6	0.242 0
Susy	0.335 0	0.325 0
KDD	0.778 0	0.775 3

表 2 给出 ESRF、使用 PCA 降维的 ESRF（ESRF-PCA）和使用 IG 降维的 ESRF（ESRF-IG），在不同数据集下的 *MAE* 和 *RMSE* 比较结果.

表 2　ESRF，ESRF-PCA，ESRF-IG 比较

数据集		DAS	HPC	Susy	KDD
ESRF	*MAE*	2.120	1.090	0.330	0.777
	RMSE	3.805	1.548	0.393	0.778
ESRF-PCA	*MAE*	1.090	1.030	0.340	0.785
	RMSE	2.094	1.064	0.399	0.785
ESRF-IG	*MAE*	2.000	0.820	0.320	0.670
	RMSE	3.088	0.894	0.320	0.670

当使用 PCA 降维减少具有低方差的强相关性变量时，ESRF 的错误率变高，PCA 无法有效处理这些数据特征，因为它捕获了具有较大方差特征的信息作为重要维度，但数据集一些有价值的信息具有较低的方差，它的预测结果变得很差. 使用 IG 降维的 ESRF 对所有实验数据集具有更有利的预测能力. 为了实现有效的 PBA 系统，本文系统的性能需要以最少的计算时间提供最佳的预测. 如果特征具有独立性和隔离性，则使用 IG 降维的 ESRF 可以为 PBA 系统提供更好的预测结果.

表 3 给出了各数据集下，Spark MLlib 在分布式环境下使用类似的计算平台实现 SRF[14] 和 ESRF 的性能评估. 从表 3 可以看出，基于 ESRF 的 PBA 系统具有高效的性能，能够处理高维大数据，这是因为 ESRF 通过超参数优化和降维技术增强 SRF. 本文的 PBA 系统使用 ESRF 算法通过将大量不同的数据转化为及时的见解来加快决策速度. 根据实验结果，本文的 PBA 系统可以在整个实验数据集上以最少的处理时间提供有效的性能.

表 3　SRF 和 ESRF 的 *MAE* 及处理时间比较

数据集		DAS	HPC	Susy	KDD
MAE	SRF[14]	2.330	1.550	0.340	0.800
	ESRF	1.090	0.820	0.320	0.670
处理时间/s	SRF[14]	130	120	140	125
	ESRF	70	65	75	68

3　结　　语

为了对高维大数据进行预测分析，本文提出了基于增强可伸缩随机森林（ESRF）的高效大数据预测分析系统（PBA）. 本文系统通过执行超参数优化来增强 SRF，基于 PCA 和 IG 通过对数据进行降维来改善预测分析的性能，将 HDFS 和 Apache Spark 分别用于分布式存储和并行计算以实现预测分析平台. 实验结果表明，本文的 PBA 系统可以提供出色的预测能力，而且可以在整个实验数据集中以最少的处理时间提供有效的性能. 未来的工作是重点研究针对高维大数据的实时 PBA 系统，以解决云分布式环境下大数据的实时跟踪和分析问题.

参考文献：

[1] OUSSOUS A, BENJELLOUN F Z, AIT LAHCEN A, et al. Big Data Technologies: a Survey [J]. Journal of King Saud University-Computer and Information Sciences, 2018, 30(4): 431-448.

[2] POUYANFAR S, YANG Y, CHEN S C, et al. Multimedia Big Data Analytics: A Survey [J]. ACM Computing Surveys (CSUR), 2018, 51(1): 1-34.

[3] CHOI T M, WALLACE S W, WANG Y. Big Data Analytics in Operations Management [J]. Production and Operations Management, 2018, 27(10): 1868-1883.

[4] YASSINE A, SINGH S, ALAMRI A. Mining Human Activity Patterns from Smart Home Big Data for Health Care Ap-

plications [J]. IEEE Access, 2017, 5(99): 13131-13141.
[5] 何兴高, 李蝉娟, 王瑞锦, 等. 基于信息熵的高维稀疏大数据降维算法研究 [J]. 电子科技大学学报, 2018, 47(2): 235-241.
[6] VENKATESH R, BALASUBRAMANIAN C, KALIAPPAN M. Development of Big Data Predictive Analytics Model for Disease Prediction using Machine learning Technique [J]. Journal of medical systems, 2019, 43(8): 272-279.
[7] WANG Y, KUNG L A, WANG W Y C, et al. An Integrated Big Data Analytics-Enabled Transformation Model: Application to Health Care [J]. Information & Management, 2018, 55(1): 64-79.
[8] WAMBA S F, GUNASEKARAN A, DUBEY R, et al. Big Data Analytics in Operations and Supply Chain Management [J]. Annals of Operations Research, 2018, 270(1): 1-4.
[9] NURAL M V, PENG H, MILLER J A. Using Meta-Learning for Model Type Selection in Predictive Big Data Analytics [C]// 2017 IEEE International Conference on Big Data. Boston: IEEE, 2017.
[10] MAHMOUD H, HEGAZY A, KHAFAGY M H. An Approach for Big Data Security Based on Hadoop Distributed File System [C]//2018 International Conference on Innovative Trends in Computer Engineering (ITCE). Aswan: IEEE, 2018.
[11] LI Yun, JIANG Yongyao, GU Juan, et al. A Cloud-Based Framework for Large-Scale Log Mining Through Apache Spark and Elasticsearch [J]. Applied Sciences, 2019, 9(6): 1114-1126.
[12] CHEN J G, LI K L, TANG Z, et al. A Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment [J]. IEEE Transactions on Parallel and Distributed Systems, 2017, 28(4): 919-933.
[13] LAKSHMIPADMAJA D, VISHNUVARDHAN B. Classification Performance Improvement Using Random Subset Feature Selection Algorithm for Data Mining [J]. Big Data Research, 2018, 100(12): 1-12.
[14] LULLI A, ONETO L, ANGUITA D. Mining Big Data with Random Forests [J]. Cognitive Computation, 2019, 11(2): 294-316.

High-Dimensional Big Data Prediction and Analysis System Based on Enhanced Scalable Random Forest

LI Fa-ling, PENG Juan

College of Software, Chongqing Institute of Engineering, Chongqing 400056, China

Abstract: To solve the program that big data is difficult to predict and to analyze because of data complexity, heterogeneity, security, scalability and large-scale data, a high-dimensional big data predictive analysis system based on enhanced scalable random forest (ESRF) has been proposed in this paper. The system enhances SRF by performing hyperparametric optimization on the training dataset, and then applies principal component analysis (PCA) and information gain (IG) to the preprocessed data to reduce the features that don't affect the model in order to reduce the processing time in the model development phase. Experimental results show that the proposed system can provide excellent prediction ability and effective performance in the whole experimental dataset with the least processing time.

Key words: high-dimensional big data; enhance scalable random forest; dimension reduction; predictive analysis; super parameter optimization

责任编辑 夏 娟