

【学术快报 / Letters】

基于统计感知的大数据系统计算框架

魏丞昊，黄哲学，何玉林

深圳大学计算机与软件学院大数据技术与应用研究所，广东深圳 518060

摘要：为在一定计算资源条件下实现大数据可计算化，本研究提出一种基于统计感知思想的 Tbyte 级大数据系统计算框架 Bigdata- α ，该框架的核心为大数据随机样本划分模型和逼近式集成学习模型。前者保证了划分后每个子数据块所包含的样本与大数据总体概率分布的一致性。后者通过分析若干个随机样本数据块替代了 Tbyte 级全量数据分析。使用 1 Tbyte 模拟数据集验证随机样本划分模型的有效性，通过逐渐增加随机样本块的个数，提升了 Higgs 数据集基分类器的分类准确度，证明该方法能克服大数据分析中计算资源的限制瓶颈。

关键词：计算机系统结构；大数据；随机样本划分；逼近式集成学习；并行分布式计算；分布式处理系统

中图分类号：TP 311 文献标志码：A doi: 10.3724/SP.J.1249.2018.05441

A statistical aware based big data system computing framework

WEI Chenghao, HUANG Zhexue, and HE Yulin

College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, Guangdong Province, P. R. China

Abstract: In order to realize the computability of big data in a certain computing resource, a statistical aware based big data system computing framework (abbreviated as Bigdata- α) is proposed in this paper to deal with Tbyte grade big data. The core of the framework is random sample partition model and asymptotic ensemble learning model. The first one guarantees the consistent distributions between the big data and its data-blocks, while the second one provides an unbiased and convergent learning model by using some samples of the big date. The effectiveness of the random sample partitioning model is verified by using a 1 Tbyte simulation dataset. By gradually increasing the number of random sample blocks, the classification accuracy of the base classifier is improved. The massive computing resources is avoided in big data analysis.

Key words: computer system structure; big data; random sample partition; asymmetric ensemble learning; paralleled distributed computing; distributed processing system

大数据分析的重要挑战之一是如何在一定的计算资源条件下、在可接受时间范围内实现大数据的可计算化^[1]。分而治之是处理大数据计算的主要策

略，即通过将大数据划分为若干小数据块文件分布式存储在集群节点上。在对大数据分析时，通过融合所有数据块并行分析结果来达到对全量大数据挖

Received: 2018-07-23; **Accepted:** 2018-08-01

Foundation: National Natural Science Foundation of China (61503252, 61473194); National Key R & D Program of China (2017YFC0822604-2); Scientific Research Foundation of Shenzhen University for Newly-Introduced Teachers (2018060)

Corresponding author: Professor HUANG Zhexue. E-mail: zx.huang@szu.edu.cn

Citation: WEI Chenghao, HUANG Zhexue, HE Yulin. A statistical aware based big data system computing framework [J]. Journal of Shenzhen University Science and Engineering, 2018, 35(5): 441-443. (in Chinese)



掘和学习的目的^[2]. Hadoop 分布式文件系统 (Hadoop distributed file system, HDFS)^[3] 主要实现大数据的划分存储和数据块文件的管理. Spark 采用弹性分布式数据集 (resilient distributed datasets, RDD) 内存数据结构将大数据分布式读入节点内存中计算, 避免了 MapReduce^[4] 反复地读写磁盘, 极大地提高了算法的运行效率^[5]. 但当数据量超出集群的最大内存容量时, Spark 算法的执行效率将大大降低, 甚至无法运行^[6-7]. 因此, 内存资源成为 Tbyte 级以上大数据的深度分析、挖掘和建模的瓶颈.

通过合适数据划分的方法, 使得大数据分布式存储的数据块可直接作为全量数据的随机样本来使用, 减少了大数据分析与建模对内存的约束. 但是, 当前的分布式文件系统的数据块文件不能被当作大数据的随机样本使用, 因此本研究在图 1 的 Bigdata- α 系统中, 提出基于随机样本划分的分布式存储模型和基于逼近式的学习框架.

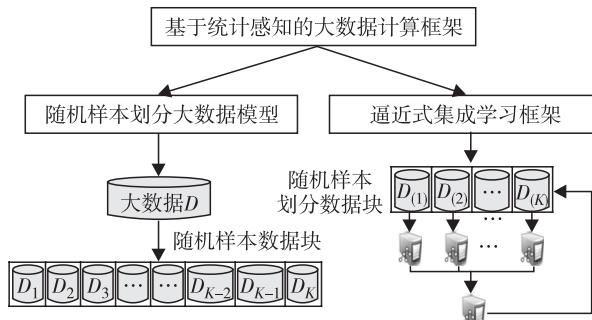


图 1 Bigdata- α 框架

Fig. 1 Bigdata- α framework

1 基于随机样本划分的大数据管理

定义 1 随机样本块. 若 $D = \{x_1, x_2, \dots, x_n\}$ 是一个大数据集的样本集合, $F(x)$ 是 D 的分布函数 (sample distribution function). 设 T 完成对 D 的某种任意划分, 则有 $T = \{D_1, D_2, \dots, D_K\}$, 若 T 中有 $E(F_k(x)) = F(x), k = 1, 2, \dots, K$ (1) 其中, $F(x)$ 为 D_k 的样本分布函数; $E(F_k(x))$ 表示其期望值. 满足这样条件下的 D_k 是 D 的一个随机样本块. 生成随机样本块的算法实现程序的代码请扫描文后二维码.

2 基于逼近式集成学习的大数据分析

利用逼近式集成学习的思想构建针对大数据的

无偏和收敛学习模型. 对于这些由以上算法得到的独立同分布数据块进行无放回抽样训练模型. 本研究以分类为例来阐述该框架, 假设大数据 D 的随机样本划分出的数据块存储在含有 R 个计算节点的集群上. 在每个计算节点上无放回抽取 τ 个随机样本块, τ 的选取以保证单个计算节点能够有效地处理当前随机样本划分数据块为准. 共训练 τR 个基分类器, 其中, 第 r 个 ($r \in \{1, 2, \dots, R\}$) 计算节点上的基分类器为 $\{C_1, C_2, \dots, C_{\tau R}\}$. 构建当前集成学习模型 $H^{(0)} = \bigcup_{r=1}^R H_r^{(0)} = \bigcup_{r=1}^R \bigcup_{t=1}^{\tau} C_n^{(0)}$, 基于独立的测试数据来验证 $H^{(0)}$ 的分类精度. 若达到设定阈值, 则停止训练; 否则, 继续在 R 个计算节点上无放回抽取随机样本数据块, 构建逼近集成学习模型 $H^{(1)} = H^{(0)} \cup \bigcup_{r=1}^R H_r^{(1)} = H^{(0)} \cup \bigcup_{r=1}^R \bigcup_{t=1}^{\tau} C_n^{(1)}$, 并验证其分类精度. 重复上述过程, 直至集成学习模型精度达到设定阈值. 逼近式集成学习框架降低了大数据分析对内存的依赖. 由于基分类器的训练是基于对大数据随机样本划分数据块的无放回抽样的数据集合完成的, 同时, 在分布式计算系统的不同计算节点上可以训练异构的基分类器, 因此有利于保证基模型之间的多样性.

3 实验结果

本研究实验使用包含 50 个计算节点的集群, 每个节点配置为 24 核 CPU、128 Gbyte 内存和 12.5 Tbyte 外存, HDFS 最大支持 128 Mbyte 的切分数据块. 图 2 为基于 Spark 模拟正态分布的 1 Tbyte 数据集 (1 亿条样本记录、每条记录包含 100 个特征值) 被划分成 1 万个随机样本数据块后相应的样本分布情况. 图中显示了随机挑选 4 个数据块的样本分布与总体样本分布具有一致性, 证明了本研究理论的有效性. 正是这种一致性保障了通过随机样本数据块所构建的大数据学习模型的无偏性和收敛性.

图 3 给出了逼近式集成学习框架利用一致分布数据块对 Higgs 数据集分类模型优化的过程. 由图 3 可见, 随着数据块的增加, 集成学习模型的精度会逐渐收敛, 其中虚线为基于大数据总体学习到的单个模型的分类精度. 当分析数据量达到整体的 15% 后, 分析精度基本与整体数据分析持平的同时, 数据量显示只需使用 10% 的数据分析结果就可达到使用 90% 数据量的整体分析精度.

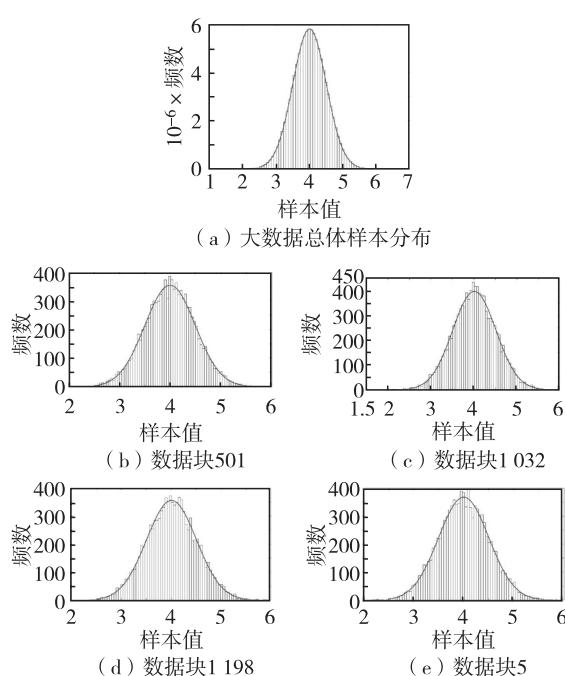


图2 大数据划分后一个样本特征分布

Fig. 2 Sample distribution after big data partition

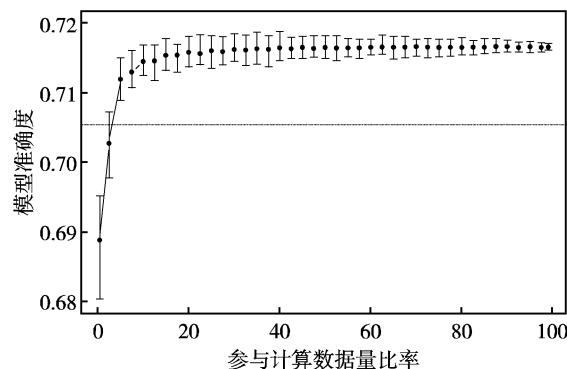


图3 逼近式集成分类模型准确度

Fig. 3 Accuracy of asymptotic ensemble classification

结语

以大数据“随机样本划分”思想为出发点, 提出基于统计感知的 Tbyte 级大数据系统计算框架 Bigdata- α 。不同于现有的 Hadoop 分布式文件系统 HDFS, Bigdata- α 保证了划分数据块与大数据总体分布的一致性。尽管单个子块存在统计偏差, 但是多个子块的集成会逼近原始数据集的统计特性最终实现了对 Tbit 级大数据的无偏收敛学习。Bigdata- α 实现了大数据在计算、统计、优化及应用 4 方面的统一。Bigdata- α 具有处理流式数据的能力, 对于与大数据总体分布一致的新增数据, 将其直接视为大

数据的随机样本数据块; 对于分布不一致的新增数据, 对其进行随机打乱, 重构新的大数据随机样本划分。

致谢: 衷心感谢深圳大学张晓亮博士的耐心指导。

基金项目: 国家自然科学基金资助项目(61503252, 61473194); 国家重点研发计划资助项目(2017YFC0822604-2); 深圳大学新引进教师科研启动资助项目(2018060)

作者简介: 魏丞昊(1986—), 男, 深圳大学博士后研究人员。研究方向: 机器学习与数据挖掘。E-mail: chenghao.wei@szu.edu.cn

引文: 魏丞昊, 黄哲学, 何玉林. 基于统计感知的大数据系统计算框架[J]. 深圳大学学报理工版, 2018, 35(5): 441-443.

参考文献 / References:

- [1] FAN Jianqiang, HAN Fang, LIU Han. Challenges of big data analysis [J]. National Science Review, 2014, 1(2): 293-314.
- [2] AHMAD A, PAUL A, RATHORE M M. An efficient divide-and-conquer approach for big data analytics in machine-to-machine communication [J]. Neurocomputing, 2016, 174(86): 439-453.
- [3] SHVACHKO H, KUANG S, RADIA W, et al. The Hadoop distributed file system [C]// Proceedings of 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies. Washington D C: IEEE Computer Society, 2010: 1-10.
- [4] DEAN J, GHEMAWAT S. MapReduce: simplified data processing on large clusters [C]// Proceedings of the 6th Conference on Symposium on Operating Systems Design & Implementation. San Francisco, USA: ACM Communication, 2008, 51(1): 107-113.
- [5] SALLOUM S, DAUTOV R, CHEN X J, et al. Big data analytics on Apache Spark [J]. International Journal of Data Science and Analytics, 2016, 1(3/4): 145-164.
- [6] SALLOUM S, Huang J Z X, HE Yulin. Empirical analysis of asymptotic ensemble learning for big data [C]// Proceedings of 2016 IEEE/ACM the 3rd International Conference on Big Data Computing, Applications and Technologies. Shanghai, China: IEEE, 2016: 8-17.
- [7] WEI C H, SALMAN S, TAMER Z E, et al. A two-stage data processing algorithm to generate random sample partitions for big data analysis [C]// International Conference on Cloud Computing. Seattle, USA: Springer, 2018: 347-364.

【中文责编: 英子; 英文责编: 子兰】



论文补充材料 团队介绍