ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 34,No. 3,May 2019,pp. 373－385
DOI:10. 16337/j. 1004-9037. 2019. 03. 001

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

大数据随机样本划分模型及相关分析计算技术

黄哲学[1,2]　何玉林[1,2]　魏丞昊[1,2]　张晓亮[1,2]

(1. 深圳大学计算机与软件学院大数据技术与应用研究所,深圳,518060;2. 深圳大学大数据系统计算技术国家工程实验室,深圳,518060)

摘　要: 设计了一种新的适用于大数据的管理和分析模型——大数据随机样本划分(Random sample partition,RSP)模型,它是将大数据文件表达成一系列RSP数据块文件的集合,分布存储在集群节点上。RSP的生成操作使每个RSP数据块的分布与大数据的分布保持统计意义上的一致,因此,每个RSP数据块是大数据的一个随机样本数据,可以用来估计大数据的统计特征,或建立大数据的分类和回归模型。基于RSP模型,大数据的分析任务可以通过对RSP数据块的分析来完成,不需要对整个大数据进行计算,极大地减少了计算量,降低了对计算资源的要求,提高了集群系统的计算能力和扩展能力。本文首先给出RSP模型的定义、理论基础和生成方法;然后介绍基于RSP数据块的渐近式集成学习Alpha计算框架;之后讨论基于RSP模型和Alpha框架的大数据分析相关计算技术,包括:数据探索与清洗、概率密度函数估计、有监督子空间学习、半监督集成学习、聚类集成和异常点检测;最后讨论RSP模型在分而治之大数据分析和抽样方法上的创新,以及RSP模型和Alpha计算框架实现大规模数据分析的优势。

关键词: 大数据;随机样本划分;渐近式集成学习;人工智能

中图分类号: TN911.73　　**文献标志码:** A

Random Sample Partition Data Model and Related Technologies for Big Data Analysis

Huang Zhexue[1,2], He Yulin[1,2], Wei Chenghao[1,2], Zhang Xiaoliang[1,2]

(1. Big Data Institute, College of Computer Science & Software Engineering, Shenzhen University, Shenzhen, 518060, China; 2. National Engineering Laboratory for Big Data System Computing Technology, Shenzhen, 518060, China)

Abstract: Random sample partition (RSP) data model distributedly represents a big data set as a set of RSP data blocks stored on a computing cluster. The RSP data model guarantees that the probability distribution of each data block is statistically consistent to the probability distribution of whole big data set. Thus, each RSP data block is a random sample of big data set and can be used to estimate the statistical properties of big data set or establish the classification and regression models. Based on the RSP data model, the big data analysis can be conducted by analyzing RSP data blocks rather than the whole big data set. This significantly reduces the computational complexity and improves the computing performance of cluster system on big data analysis. In this paper, we firstly present the definition, basic theory and generation method of RSP. Second, we introduce an asymptotic ensemble learning framework

基金项目: 国家重点研发计划(2017YFC0822604-2)资助项目;中国博士后科学基金(2016T90799)资助项目;深圳大学2018年度新引进教师科研启动基金(2018060)资助项目;广东省普通高校国家级重大培育基金(2014GKXM054)资助项目。
收稿日期: 2018-08-23;**修订日期:** 2019-03-01

called Alpha framework used for big data analysis. Third, we discuss the main big data analysis methods based on the RSP data model and Alpha framework, including data exploration & cleaning, probability density function estimation, supervised subspace learning, semi-supervised ensemble learning, clustering ensemble and outlier detection. Finally, we discuss the innovations and advantages of the RSP data model and Alpha framework in big data analysis by using the divide-and-conquer strategy on random samples.

Key words: big data; random sample partition; asymptotic ensemble learning; artificial intelligence

引　言

数据分析是挖掘大数据价值的重要手段和途径。数据文件通常被表示成个对象(或记录)的集合$\boldsymbol{D}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$,其中每个对象被表示成$M$个属性(或特征)的向量$\boldsymbol{x}_n=(x_{n1},x_{n2},\cdots,x_{nM}),n=1,2,\cdots$(当$M$很大时,$\boldsymbol{D}$为超高维数据;当$N$很大时,$\boldsymbol{D}$为大数据;当$M$和$N$都很大时,$\boldsymbol{D}$为超高维大数据)。正如陈国良院士[1]描述的那样:“到目前为止尚没有这样一个被普遍认可的大数据定义出现”,仅能够从大数据的特征对其展开描述,其中比较有代表性的是大数据的“4V”特征。给定数据文件或数据集$\boldsymbol{D}$,数据分析的任务主要包括:数据预处理、数据探索分析、奇异值检测、关联分析、聚类分析、分类和预测等。

目前可用于数据分析的方法和算法很多,计算复杂度高是许多数据分析和机器学习算法的共有特点,因此,这些算法很难用于大数据分析。分而治之是当前大数据分布式并行处理普遍采用的策略,其步骤是将大数据文件$\boldsymbol{D}$切分成个小的数据块文件,分布式地存储在集群节点上。对大数据分析时,先在每个节点上对小数据块进行计算,然后把小数据块的计算结果传送到主节点进行综合分析,得到大数据的分析结果,对于复杂的分析算法,上述两个步骤需要迭代进行,算法需要按分而治之的计算模式并行实现。当前主流的大数据处理平台Hadoop MapReduce和Apache Spark采用的都是分而治之策略。

大数据的划分和数据块文件的管理采用分布式文件系统,如HDFS[2-3];而大数据的分析算法则采用MapReduce[4-5]或Spark[6]计算模型实现。近年来,国内对于MapReduce和Spark的研究主要集中在提高它们解决大数据分析问题的效率上。2017年,张滨[7]对MapReduce大数据并行处理过程中的查询优化控制、数据分布优化和调度优化控制等若干关键技术进行了研究。2018年,李志斌[8]通过引入基于内存的PageRank算法设计了一个针对大规模图数据集的MapReduce大数据并行处理系统。2018年,王晨曦等人[9]提出了面向大数据处理的基于Spark的异质内存编程框架,解决了如何将数据合理地布局到异质内存的问题。2019年,宋泊东等人[10]通过使用Apache Kafka作为消息中间件设计了一种基于Spark的分布式大数据分析算法。更多关于MapReduce和Spark框架性能分析的介绍可参见吴信东等人的研究工作[11]。

基于MapReduce的分布式计算框架通过磁盘文件进行Map节点与Reduce节点之间的数据交换,在运行循环迭代算法时需要大量的磁盘读写操作,极大地降低了算法的执行效率。Spark采用RDD内存数据结构将大数据分布式存储在节点的内存中,在运行迭代式分而治之的算法时不需要反复地读写磁盘,在一定程度上提升了算法的运行速度。但是,当数据量超出内存容量时,Spark的算法执行效率将被大大降低,甚至无法运行。因此,内存资源成为TB级以上大数据的深度分析、挖掘和建模技术的瓶颈。

采用随机样本数据对大数据做统计估计和建模是大数据分析的有效途径。但是,对分布式大数据

做随机抽样是计算成本很高的操作,因为要对大数据的分布式数据块文件进行遍历才能获得随机样本数据,这一过程需要大量的磁盘读写操作和节点间的数据通信。如果大数据的分布式数据块可以直接当作样本数据来用,大数据的随机抽样操作就不需要了,有了可用的样本数据,大数据的分析与建模问题就可以通过对样本数据的分析与建模解决,这样就减少了大数据分析对内存的约束。但是,当前的大数据分布式数据块,即HDFS数据块文件不能当作大数据的随机样本使用,因为不同数据块的数据分布不同,数据块的数据分布与大数据本身的数据分布也不同,简单地将数据块当作大数据的随机样本数据使用,会产生统计意义上不正确的结果。

针对当前大数据分析的技术瓶颈,本文介绍一个新的将统计方法与集群计算融合的大数据分析解决方案。针对分布式大数据抽样的问题,本文提出大数据随机样本划分(Random sample partition,RSP)模型来表达分布式大数据。RSP模型同样将大数据划分成小的数据块分布式存储在集群的节点上,但每个数据块的样本数据分布与整个大数据的样本数据分布保持一致,这样就可以将存储在节点上的数据块文件直接拿来当作随机样本数据使用,采用统计中普遍使用的样本估计方法估计大数据的统计量,采用机器学习中的集成学习方法建立大数据集成学习模型。基于RSP数据块的大数据分析不需要对整个大数据进行计算,极大地降低对内存的需求,具有更大的数据扩展性,突破了TB级大数据的计算技术瓶颈。

本文首先介绍大数据随机样本划分模型的定义、存在性定理和大数据随机样本划分的生成算法。然后介绍了基于RSP数据表达模型的大数据渐近式集成学习框架——Alpha框架以及基于此框架和RSP数据块的大数据分析方法,包括大数据探索与清洗、大数据概率密度函数估计、大数据子空间学习、大数据半监督集成学习、大数据聚类集成和大数据异常点检测等。最后总结采用RSP数据模型和Alpha框架进行大数据分析的优越性和创新性。

1 随机样本划分模型

随机样本划分模型的核心思想是将大数据文件划分成许多小的随机样本划分数据块文件,即每个数据块文件是大数据文件的一个随机样本数据。这样的划分给大数据分析带来两个好处:(1)随机样本数据可以直接通过选择数据块文件获得,不需要对大数据的单个记录进行抽样,避免了分布式大数据随机抽样的操作;(2)通过对少量数据块文件的分析和建模即可得到大数据的统计估计结果和模型。采用随机样本划分模型,大数据分析的工作转变成对随机样本数据块文件的分析与建模,极大地减少了大数据分析的计算量,提高了大数据分析的能力。本节对大数据随机样本划分的理论基础和大数据随机样本数据块的生成方法进行详细阐述。

1.1 RSP模型的理论基础

在定义随机样本划分之前,本文首先定义大数据划分。

定义1(大数据划分) 设T是由操作T生成的大数据的一组子集$D_1,D_2,\cdots,D_k$构成的集合,即$T=\{D_1,D_2,\cdots,D_k\}$,如果T满足以下两个条件,则称T是$\boldsymbol{D}$的一个划分:(1)对于任意的$i,j\in\{1,2,\cdots,k\}$且$i\neq j$,$D_i\cap D_j=\varnothing$;(2)$\bigcup_{k=1}^{k}D_k=\boldsymbol{D}$,同时称$T$是大数据$\boldsymbol{D}$的一个划分操作。

由定义1可知,在HDFS分布式文件系统中,大数据表达成数据块文件的划分,HDFS数据块文件被分布式存储在集群节点上。在一般情况下,HDFS数据块文件不能作为大数据的随机样本数据使用,因为数据块文件的数据分布与大数据的数据分布不一致。为解决分布不一致的问题,本文给出大数据随机样本划分定义[12]。

定义2(随机样本划分数据块) 设T是大数据$\boldsymbol{D}$的一个划分操作,$T=\{D_1,D_2,\cdots,D_k\}$是由T

生成的$\boldsymbol{D}$的含有k个子集的一个划分，记$\widetilde{F}(D_k)$和$F(\boldsymbol{D})$分别表示数据子集D_k和大数据$\boldsymbol{D}$的概率分布函数。对于任意$k\in\{1,2,\cdots,k\}$，如果$E[\widetilde{F}(D_k)]=F(\boldsymbol{D})$成立，$E[\widetilde{F}(D_k)]$表示分布$\widetilde{F}(D_k)$的期望，则称$\boldsymbol{T}$是$\boldsymbol{D}$的一个随机样本划分，$D_1,D_2,\cdots,D_k$是$\boldsymbol{D}$的随机样本划分数据块，简称RSP数据块。

下面给出定理1，确保对于任何大数据都可以将其表达成一组RSP数据块，本文将RSP数据块称之为大数据随机样本划分数据模型，或RSP模型。

定理1（RSP存在性定理） 设大数据$\boldsymbol{D}$有N个记录，$N_1,N_2,\cdots,N_k$，是满足$\sum_{k=1}^{K}N_k=N$的$k(k>1)$个正整数，则存在一个划分操作$\boldsymbol{T}$，使得由$\boldsymbol{T}$生成的大数据划分$T=\{D_1,D_2,\cdots,D_k\}$是$\boldsymbol{D}$的随机样本划分，其中D_k含有N_k个记录，$k\in\{1,2,\cdots,K\}$。

证明：对于任意给定的含有N个对象的大数据$\boldsymbol{D}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$，随机选取一个$N$元排列$\tau=\{\tau_1,\tau_2,\cdots,\tau_N\}$。将$\boldsymbol{D}$的全部$N$个对象按$\tau_n(n=1,2,\cdots,N)$值的大小重新排序，得到$\boldsymbol{D}'=\{\boldsymbol{x}'_1,\boldsymbol{x}'_2,\cdots,\boldsymbol{x}'_N\}$，其中$\boldsymbol{x}'_n=\boldsymbol{x}_{\tau_n}$。将$\boldsymbol{D}'$按顺序切分成$k$个子集$D_1,D_2,\cdots,D_k$，其中每个子集分别含有$N_1,N_2,\cdots,N_k$个记录。则对任意$D_k,k\in\{1,2,\cdots,K\}$以及$\boldsymbol{D}$中任意一个元素$\boldsymbol{x}_n,n\in\{1,2,\cdots,N\}$，有$P(x_n\in D_k)=\frac{N_k}{N}$成立。记$F_k(x)$和$F(x)$分别表示数据子集$D_k$和大数据$\boldsymbol{D}$的概率分布函数。对任意实数$x$，由样本分布函数的定义知，$\boldsymbol{D}$中取值不大于$x$的对象数为$N\times F(x)$，所以$D_k$中取值不大于$x$的对象数的期望为$N\times F(x)\times\frac{N_k}{N}=N_k\times F(x)$，所以$F_k(x)$的期望为$E[F_k(x)]=\frac{N_k\times F(x)}{N_k}=F(x)$。由$k$的任意性知$T=\{D_1,D_2,\cdots,D_K\}$为大数据$\boldsymbol{D}$的一个RSP。

简便起见，这里只考虑对象取值为一维时的情况。当对象取值为向量时，证明方法类似。定理1保证了对于任意大数据，本文都能通过随机样本划分操作将它转换成RSP表达。由定义2可知，每个RSP数据块的概率分布函数与大数据$\boldsymbol{D}$的概率分布函数保持一致性。但是，这种一致性是在期望意义下的，所以每个具体的RSP数据块的概率分布函数与大数据$\boldsymbol{D}$的概率分布函数不完全相同。当然，RSP数据块之间的概率分布函数相似度也有所不同。相似度越高，两个数据块之间相互表达的准确度越高。

给定一个大数据$\boldsymbol{D}$的随机样本划分T，本文采用如下公式计算两个RSP数据块的概率密度相似性和RSP数据块与$\boldsymbol{D}$的概率密度相似性。首先，如果$D_i=\{\boldsymbol{x}_1^{(i)},\boldsymbol{x}_2^{(i)},\cdots,\boldsymbol{x}_i^{(i)}\}$和$D_j=\{\boldsymbol{x}_1^{(j)},\boldsymbol{x}_2^{(j)},\cdots,\boldsymbol{x}_j^{(j)}\}$，$i,j\in\{1,2,\cdots,K\}$且$i\neq j$，满足

$$gmmd(D_i,D_j)<\sqrt{\frac{8u^2(N_i+N_j)}{N_iN_j}\log_2(\alpha)^{-1}} \tag{1}$$

则称D_i和D_j具有α显著性水平下的概率分布一致性，其中$gmmd(\cdot,\cdot)$为基于再生核希尔伯特空间核函数$kernel(\cdot,\cdot)$构造的推广最大平均差异（Generalized maximum mean discrepancy，GMMD）[13-14]，表达式为

$$gmmd(D_i,D_j)=\frac{1}{N_i(N_i-1)}\sum_{n=1}^{N_i}\sum_{m\neq n}^{N_i}kernel[\boldsymbol{x}_n^{(i)},\boldsymbol{x}_m^{(i)}]+\frac{1}{N_j(N_j-1)}\sum_{n=1}^{N_j}\sum_{m\neq n}^{N_j}kernel[\boldsymbol{x}_n^{(j)},\boldsymbol{x}_m^{(j)}]-\frac{2}{N_iN_j}\sum_{n=1}^{N_i}\sum_{m=1}^{N_j}kernel[\boldsymbol{x}_n^{(i)},\boldsymbol{x}_m^{(j)}] \tag{2}$$

式中：N_i和N_j为数据块D_i和D_j包含记录的个数，u为核函数$kernel(\cdot,\cdot)$的上界，$kernel(\cdot,\cdot)$可以选择径向

基函数核。之后,构造$\frac{K(K-1)}{2}$个数据块对(D_i,D_j),$i=1,2,\cdots,K-1$,$j=i+1,i+2,\cdots,K$,如果式(1)成立的次数大于等于$\frac{K(K-1)}{4}+Z_{\frac{\alpha}{2}}\times\frac{\sqrt{K(K-1)}}{2\sqrt{2}}$,其中$Z_{\frac{\alpha}{2}}$为正态分布的$\frac{\alpha}{2}$分位数,则将$D_1,D_2,\cdots,D_K$判定为$\alpha$显著性水平下的概率同分布数据块。在实践中,本文可以用式(1)来检验RSP的大数据表达,也可以在RSP抽样过程中通过式(1)来选择RSP数据块。

1.2 RSP模型的生成方法

定理1的证明给出了一个RSP数据模型的生成方法,但此方法需要对整个大数据进行排序,当大数据的对象数很大时,在分布式计算环境下对大数据的排序是计算量很大的操作,费时或者难于完成。

为了解决分布式环境下大数据RSP的生成问题,本文提出了分为两步的计算方法[15]:第一步先将大数据切成$P(P>1)$个较大的数据块$D_1,D_2,\cdots,D_P$,再将每个数据块按定理1证明中的方法分别随机打乱,切分成$Q(Q>1)$个更小的RSP数据块$D_{p1},D_{p2},\cdots,D_{pQ}$,$p=1,2,\cdots,P$,生成$P$个RSP集合;第二步将每个RSP中的对应RSP数据块$D_{1q},D_{2q},\cdots,D_{Pq}$,$q=1,2,\cdots,Q$,合并生成新的数据块$D'_q$,新生成的大数据划分$D'_1,D'_2,\cdots,D'_Q$是大数据的一个RSP。这一方法的正确性可以通过定理2证明。

定理2 设D_1和D_2是分别含有N_1和N_2个对象的两个数据块,$D_{1.}$和$D_{2.}$分别是D_1和D_2的含有$N_{1.}$和$N_{2.}$个对象的RSP数据块,当$\frac{N_{1.}}{N_1}=\frac{N_{2.}}{N_2}$时,$D_{1.}\cup D_{2.}$是$D_1\cup D_2$的RSP数据块。

证明:设D_1和D_2的概率分布函数分别为$F_1(x)$和$F_2(x)$,$D_{1.}$和$D_{2.}$的概率分布函数分别为$F_{1.}(x)$和$F_{2.}(x)$,则有$E[F_{1.}(x)]=F_1(x)$和$E[F_{2.}(x)]=F_2(x)$。对于任意实数x,$D_{1.}$和$D_{2.}$中取值不大于x的对象数分别为$N_{1.}\times F_{1.}(x)$和$N_{2.}\times F_{2.}(x)$,所以$D_{1.}\cup D_{2.}$的概率分布函数为$F_{1.\cup 2.}(x)=\frac{N_{1.}\times F_{1.}(x)+N_{2.}\times F_{2.}(x)}{N_{1.}+N_{2.}}$。同理可得$D_1\cup D_2$的概率分布函数为$F_{1\cup 2}(x)=\frac{N_1\times F_1(x)+N_2\times F_2(x)}{N_1+N_2}$。从而可计算$F_{1.\cup 2.}(x)$的期望为

$$E[F_{1.\cup 2.}(x)]=E\left[\frac{N_{1.}\times F_{1.}(x)+N_{2.}\times F_{2.}(x)}{N_{1.}+N_{2.}}\right]=\frac{N_{1.}\times E[F_{1.}(x)]+N_{2.}\times E[F_{2.}(x)]}{N_{1.}+N_{2.}}=$$
$$\frac{N_{1.}\times F_1(x)+N_{2.}\times F_2(x)}{N_{1.}+N_{2.}}=\frac{N_1\times F_1(x)+N_2\times F_2(x)}{N_1+N_2}=F_{1\cup 2}(x)$$

所以,$D_{1.}\cup D_{2.}$是$D_1\cup D_2$的RSP数据块。

简便起见,这里只考虑对象取值为一维时的情况。当对象取值为向量时,证明方法类似。根据定理2进行推理,可以证明合并后的每个数据块都是大数据的一个随机样本,因此,Q个数据块是大数据的一个RSP。图1展示了上述大数据RSP的两步生成算法实现过程。

图2展示一个真实数据HDFS数据块的一个属性分布和RSP数据块的同一属性分布。可以看到,RSP数据块的属性数据分布同大数据的属性数据分布相似,因此RSP数据块可以作为随机样本来分析大数据(图2(b))。HDFS数据块之间的属性数据分布不同,同时与大数据的属性数据分布也不同,因此HDFS数据块不能当作随机样本数据块来用(图2(a)),要想得到大数据的正确结果,必须对整个大数据进行分析。

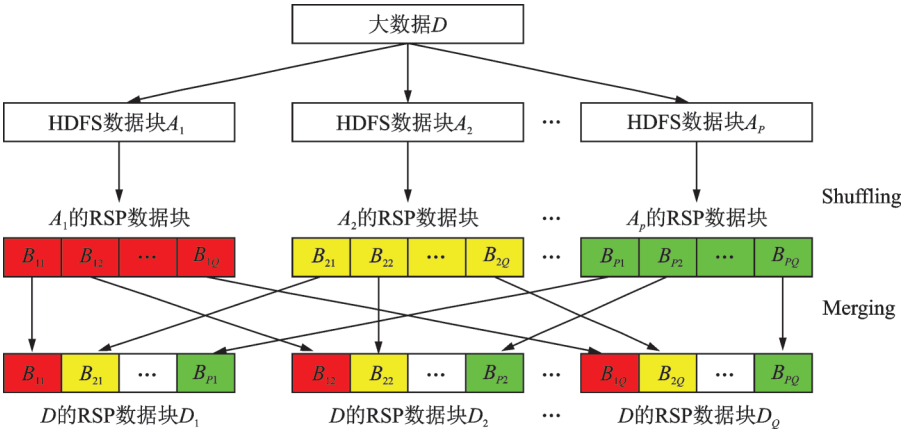
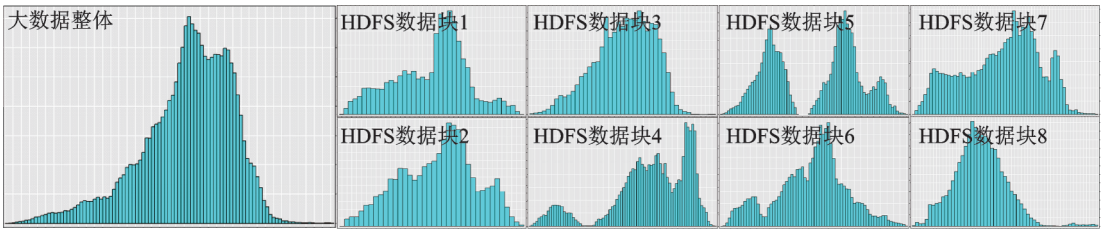


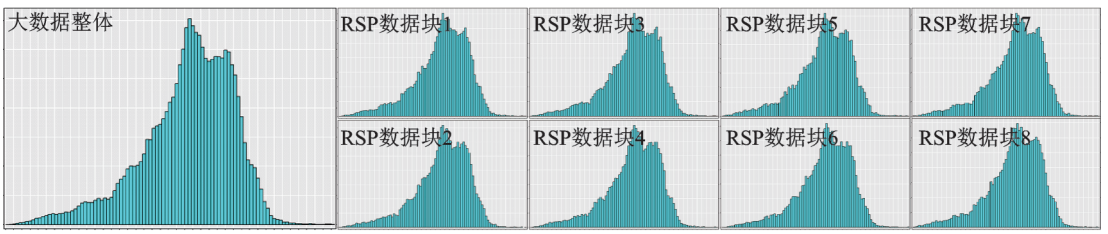
图1 大数据RSP的生成过程

Fig.1 RSP model of big data



(a) HDFS数据块与大数据整体分布不一致

(a) Inconsistent distribution between HDFS data block and whole big data



(b) RSP数据块与大数据整体分布一致

(b) Consistent distribution between RSP data block and whole big data

图2 不同数据模型中数据块与大数据整体分布一致性的对比

Fig.2 Distribution comparison between data block and whole big data in different data management models

笔者已经在Spark上实现了上述大数据RSP的两步生成算法，初步的实验结果显示，在27个计算节点的集群上，生成1 TB大数据（含100亿个对象，10属性）的RSP（30 000个RSP数据块）大概需要56 min。在实际应用中，对每个大数据的RSP生成只需要一次。

2 基于RSP模型的计算框架

第1节介绍了RSP数据模型的理论基础和在分布式环境下生成大数据RSP的算法。在分布式集群上，针对某个大数据D生成的RSP数据块随机分布存储在计算节点的磁盘上，为了有效地利用分布的RSP数据块对大数据进行分析，本文设计开发了基于RSP数据块的渐近式集成学习框架——Alpha计算框架[16-17]。下面介绍Alpha计算框架和在此框架下的大数据分析流程。

Alpha计算框架是基于RSP数据块的分布式大数据处理与分析框架,其设计思想是基于RSP模型的相关理论。处理和分析一个给定的大数据$\boldsymbol{D}$,当把$\boldsymbol{D}$转换成RSP数据块$D_1,D_2,\cdots,D_K$,并将其分布存储在计算节点磁盘上后,针对任意一个RSP数据块$D_k,k\in\{1,2,\cdots,K\}$,进行处理和分析都能得到$\boldsymbol{D}$的一个统计量θ_k的估计值,而θ_k估计值的期望值是$\boldsymbol{D}$的统计量θ值[18]。因此,θ_k是θ的近似值,存在一定的误差。如果采用多个RSP数据块的来计算θ的估计值,其估计值的误差将随着RSP数据块的增加而下降。

根据上述统计估计原理和分布式环境下分而治之的大数据处理策略,本文设计了渐近式集成学习Alpha计算框架如图3所示,本文仅以分类问题来阐述Alpha框架的工作原理。给定大数据$\boldsymbol{D}$的RSP数据块集合$\{D_1,D_2,\cdots,D_K\}$,假设集群中计算节点的个数为$Q(Q<K)$。采用无放回块抽样随机抽取个RSP数据块$\{D_1^{(1)},D_2^{(1)},\cdots,D_Q^{(1)}\}$,每个计算节点抽取一个,上标(1)表示是第一批抽样;在Q个节点上采用同一算法从Q个RSP数据块训练Q个基分类器$\{h_1^{(1)},h_2^{(1)},\cdots,h_Q^{(1)}\}$,每个基分类器在各节点上独立计算完成;在主节点上构建集成分类器H_1;采用独立样本测试集测试H_1,如果精度达到了设定的阈值条件,输出H_1,训练结束;否则,进行第二批RSP数据块无放回抽样,按相同方式训练第二批基分类器$\{h_1^{(2)},h_2^{(2)},\cdots,h_Q^{(2)}\}$,在主节点上构建集成分类器$H_2$,采用统一样本测试集测试$H_1\cup H_2$,如果精度达到了设定的阈值条件,输出$H_1\cup H_2$,否则,进行新的一次RSP抽样和建模过程。不断重复上述过程,直到$\bigcup_{j=1}^{J}H_j$满足设定的阈值条件,其中J为逼近阈值条件的迭代次数。理论推导可以得出上述渐近式集成学习Alpha框架的收敛条件为

$$P\left\{\left|\bigcup_{j=1}^{J}H_j-H\right|\geqslant\delta\right\}\leqslant\frac{K^2}{4N\delta^2(JQ)^2}\quad JQ\leqslant K \tag{3}$$

式中:H为在大数据$\boldsymbol{D}$上的学习模型(这是一个期望模型,实际不存在),N为大数据$\boldsymbol{D}$含有对象的个数,

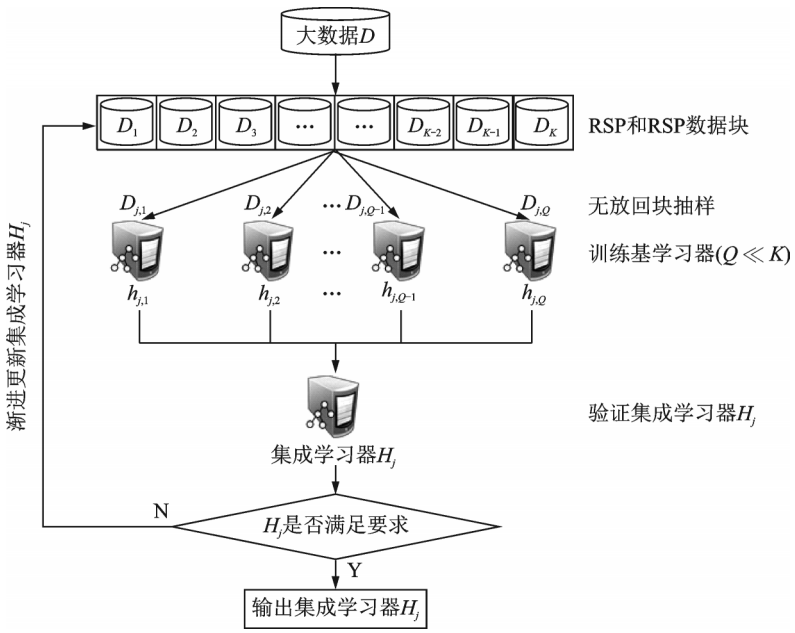


图3 大数据渐近式集成学习Alpha计算框架

Fig.3 Alpha framework for big data analysis and computation

K为RSP数据块的个数，$\delta > 0$为正数阈值。由式(3)可见，当RSP数据块的个数K和集群计算节点个数Q确定之后，$\frac{K^2}{4N\delta^2(JQ)^2}$的取值仅与逼近阈值条件的迭代次数$J$相关，当$J \to +\infty$时，$\frac{K^2}{4N\delta^2(JQ)^2} \to 0$。这表明Alpha框架能够保证学习算法的收敛性。

Alpha计算框架的样机已经在Microsoft R Server、Apache Spark和HDFS集群上实现。图4展示真实数据HIGGS上4个特征的均值和标准差的渐近估计值，每一个子图的横轴代表使用数据的数据量占数据总量的百分比。图5展示HIGGS数据基于RSP数据块的渐近集成分类模型精度与所有数据建模的精度对比，每一个子图的横轴代表使用数据的数据量占数据总量的百分比，其中60个RSP数据块，每个数据块大约183 753条记录；100个RSP数据块，每个数据块110 000条记录；200个RSP数据块，每个数据块55 000条记录。本文从图5中可以发现只用少量的RSP数据块就可以达到从全部数据学习的单一模型的精度，该实验证实了式(3)的合理性。图6是集成学习模型的计算时间与单个模型计算时间的对比，其中图6(b)最右侧的灰色柱代表单个模型计算时间。本文从图6中可以发现基于RSP数据块的Alpha计算框架极大地提高了集群计算的效率。

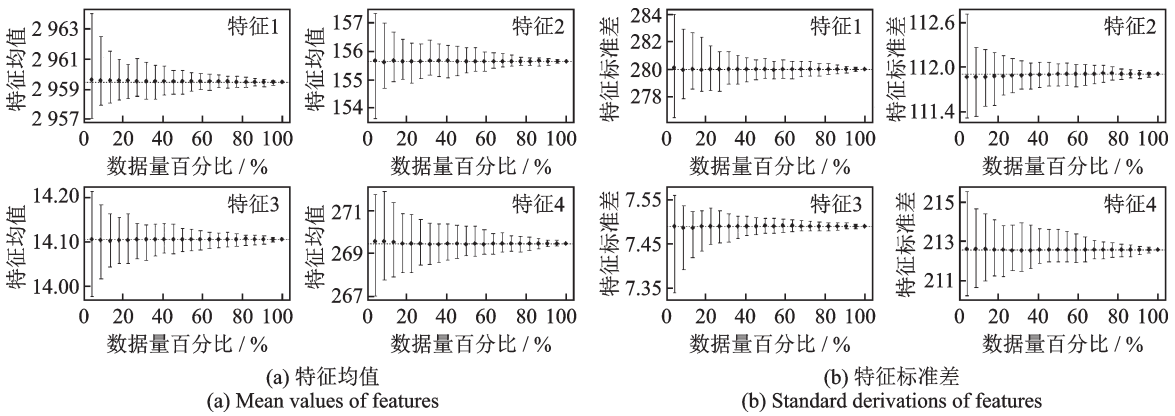


图4 HIGGS数据4个特征均值和标准差的渐近估计值

Fig.4 Feature approximations of mean and standard derivation in HIGGS data set

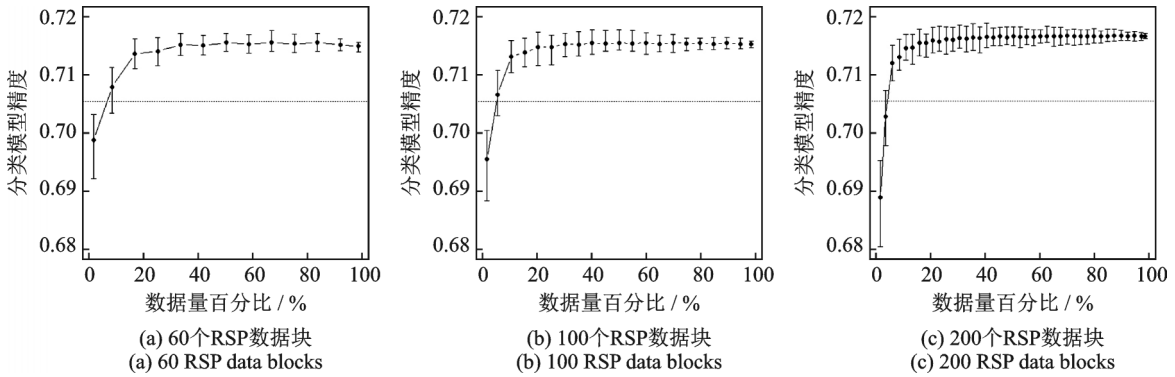


图5 基于RSP数据块的集成分类模型与基于所有数据建模的精度对比

Fig.5 Accuracy comparison between asymptotic ensemble learning model and single learning model

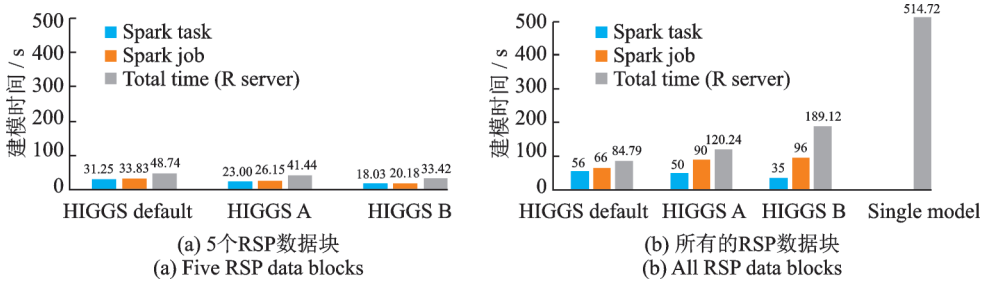


图6 渐近式集成学习模型与单个模型计算时间的对比

Fig.6 Time comparison between asymptotic ensemble learning model and single learning model

3 基于RSP模型的大数据分析技术

基于RSP模型和Alpha计算框架可以设计开发一系列新的分布式大数据分析方法和技术,这些分析方法的核心思想是:根据分而治之的策略,在分布式集群系统上利用Alpha计算框架,随机抽取RSP数据块的子集计算大数据的统计量估计值,建立大数据的集成模型。本节介绍以下6种基于RSP数据模型和Alpha计算框架的大数据分析方法。

3.1 数据探索和清洗

数据探索是数据分析的重要步骤,分析一个未知的大数据D,首先需要做的工作是要理解D,要知道D的各属性的分布,了解D的各种数据错误,设计数据清洗流程对D进行清理,改正数据错误。有了D的RSP数据模型后,数据探索和清洗可以在RSP数据块上进行,可以极大地降低工作量,提高数据理解和清洗的效率。

因为RSP数据块具有同整个大数据D一致的分布,可以随机抽取一个或几个RSP数据块,用可视化工具展示数据块属性的分布,通过数据块属性的分布即可理解大数据的属性分布,其原理如图2(b)所示。同理,可以通过处理RSP数据块找出数据的错误,设计清洗错误的过程,再将清洗过程应用在其他RSP数据块。由于错误数据重复出现在RSP数据块中的概率大致相同,在少量随机抽取的RSP数据块中发现的错误数据反映了大数据中的主要错误数据,通过Alpha计算框架的多次迭代,即可发现大数据D中的大部分错误数据。由于错误数据的发现过程是在RSP数据块上进行的,要比从整个大数据上发现错误的效率高。

3.2 概率密度函数估计

概率密度函数(Probability density function,PDF)是随机变量统计特性的集中体现,估计概率密度分布是数据分析的重要任务,也是大数据分析的一大挑战。RSP数据模型提供了一种“局部推断整体”的间接式估计大数据PDF的途径。

假设大数据D的属性变量均为连续值,对应的PDF为$\tilde{f}$,RSD数据块$D_1,D_2,\cdots,D_K$对应的PDF分别为$\tilde{f}_1,\tilde{f}_2,\cdots,\tilde{f}_K$,其中$\tilde{f}_k$,$k=1,2,\cdots,K$可以通过核密度估计方法求得。直观的想法,由于$D_1,D_2,\cdots,D_K$与$D$存在概率分布上的一致性,可以通过建立式(4)确定$\tilde{f}$为

$$\tilde{f}\approx\frac{1}{Q}\sum_{q=1}^{Q}\tilde{f}_q \tag{4}$$

式中$Q(Q<K)$是随机抽取的RSP数据块个数。即在α显著性水平下大数据D的PDF可由RSP数据块PDF的均值表示。可以看出,$\tilde{f}$可以采用Alpha计算框架获得,通过将$\tilde{f}_q$,$q=1,2,\cdots,Q$在集群节点上独立地算出。

3.3 有监督子空间学习

采用RSP样本数据建立分类或回归集成模型，由于RSP数据块分布的一致性，降低了基分类器的多样性，影响集成模型的精度。为增加基分类器的多样性，可以采用子空间抽样的方法，对RSP数据块抽取不同的属性子集来学习基分类器。给定大数据 $\boldsymbol{D}$ 的属性变量集合 $\{A_1,A_2,\cdots,A_M\}$ 和 Q 个RSP数据块集合 $\{D_1,D_2,\cdots,D_Q\}$，为每个RSP数据块随机抽取一个 L 维的子空间，得到 Q 个不同的子空间 $\{A_1^{(q)},A_2^{(q)},\cdots,A_L^{(q)}\}\subset\{A_1,A_2,\cdots,A_M\}$，$q\in\{1,2,\cdots,Q\}$，$L<M$。

根据Alpha计算框架，子空间抽样可以在节点上的RSP数据块独立进行，对每个RSP数据块的子空间数据进行独立建模，生成基分类器 $\{h'_1,h'_2,\cdots,h'_Q\}$，最后获得集成模型

$$H=\bigcup_{q=1}^{Q}h'_q \tag{5}$$

由于每个基模型 h'_q，$q\in\{1,2,\cdots,Q\}$ 从不同子空间数据得来，因此增加了基模型的多样性，提高渐近集成学习模型的性能。

3.4 半监督集成学习

对于含有少量有标签数据和大量无标签数据的大数据，采用RSP数据模型在Alpha计算框架下最大限度地利用无标签数据提升基于有标签数据训练的集成模型的泛化能力。半监督集成学习[19]是一种融合了半监督学习和集成学习优势的学习方法，基于RSP数据模型，在Alpha计算框架下可以设计如下集成学习算法：

(1)随机抽取 Q 个RSP数据块 $\{D_1,D_2,\cdots,D_Q\}$，将这些数据块中的有标签数据抽取出来合并成一个训练数据集 $\boldsymbol{D}_T$。

(2)对 $\boldsymbol{D}_T$ 做 Q 次放回抽样生成 Q 个训练数据集，放回 $\{D_1,D_2,\cdots,D_Q\}$ 的相应节点，根据Alpha框架训练个基模型 $\{h_1^{(0)},h_2^{(0)},\cdots,h_Q^{(0)}\}$，并构建集成模型 $H^{(0)}=\bigcup_{q=1}^{Q}h_q^{(0)}$。

(3)使用 $H^{(0)}$ 对 $\{D_1,D_2,\cdots,D_Q\}$ 中的无标签数据进行打标。

(4)将打标的数据与同节点的相应训练数据合并，重新训练一组基模型 $\{h_1^{(1)},h_2^{(1)},\cdots,h_Q^{(1)}\}$，并构建集成模型 $H^{(1)}=\bigcup_{q=1}^{Q}h_q^{(1)}$。再用 $H^{(1)}$ 对 $\{D_1,D_2,\cdots,D_Q\}$ 中无标签数据进行打标。不断重复上述过程，直至集成模型表现稳定。

获得稳定的模型 H 后，用它对其他没有选择的RSP数据块中无类标的数据进行打标。如果有类标的数据极少，可以抽取大数据所有有类标的数据做有放回抽样。

3.5 聚类集成

基于大数据 $\boldsymbol{D}$ 的RSP数据块 $\{D_1,D_2,\cdots,D_K\}$ 的聚类集成是一大挑战，因为已有的聚类集成方法都是集成同一数据集的不同聚类结果，而被聚类的对象是同一对象集合[20]，在集成不同聚类结果时，有相同的对象标识可以参考。而不同的RSP数据块包含不同的对象集合，在集成不同RSP数据块聚类结果时没有相同的标识可以参考。因此，需要采用不同RSP数据块的簇的统计特征集成聚类结果。

假设 $\{C_1^{(i)},C_2^{(i)},\cdots,C_R^{(i)}\}$ 和 $\{C_1^{(j)},C_2^{(j)},\cdots,C_P^{(j)}\}$ 是从RSP数据块 D_i 和 D_j，$i,j\in\{1,2,\cdots,K\}$ 且 $i\neq j$，分别得到的两组簇的集合，可以采用一个相似性的度量 $s[C_r^{(i)},C_p^{(j)}]$，$r\in\{1,2,\cdots,R\}$，$p\in\{1,2,\cdots,P\}$，计算不同数据块的簇之间的相似性，根据相似性对不同数据块的簇进行合并，合并后重新计算新的簇的特征量，如簇的中心点，再根据新的中心点对其他RSP数据块进行聚类。

基于RSP数据块的聚类集成过程可以采用不同的聚类算法[21-22]生成每个RSP数据块的聚类簇,可以采用不同的簇之间的相似性度量,可以采用不同的簇合并方法进行簇的合并,整个过程可以在Alpha计算框架上完成。

3.6 异常点检测

异常点检测[23]是数据分析的一个重要任务,在许多领域有应用需求,大数据中的异常点检测也是当前的一大挑战,RSP数据模型提供了异常点检测的新途径。

基于RSP数据块的异常点检测通过两步进行:第一步是从随机选出的Q个RSP数据块$\{D_1,D_2,\cdots,D_Q\}$中发现异常点,已有的异常点检查算法都可以在这一步采用;第二步是判定从每个RSP数据块发现的异常点是否是大数据的异常点。假设数据$\boldsymbol{x}_*^{(q)}\in D_q,q\in\{1,2,\cdots,Q\}$,是RSP数据块$D_q$的异常点,可以用下面两种方法判断$\boldsymbol{x}_*^{(q)}$是否为大数据$\boldsymbol{D}$的异常点。

(1)基于数据块的概率密度函数判定法:对于给定的阈值$\varepsilon>0$和显著性水平α,检验$\bar{f}_p[\boldsymbol{x}_*^{(q)}]>\varepsilon$,$p=1,2,\cdots,Q$且$p\neq q$成立的次数,如果未达到某种假设检验的标准,则认为$\boldsymbol{x}_*^{(q)}$是大数据的异常点。

(2)基于数据信息量的判定法:即将RSP数据块D_q的异常点$\boldsymbol{x}_*^{(q)}$加到其他RSP数据块上会否引起信息量的激增。通常情况下,非异常数据的增加不会引起数据集信息量的激增。如果$\boldsymbol{x}_*^{(q)}$未能够引起大多数RSP数据块信息量的激增,那么则认为$\boldsymbol{x}_*^{(q)}$是大数据的异常点。

4 结束语

RSP数据模型将大数据划分成随机样本数据块文件分布式存储,由于任何一个RSP数据块的分布都与大数据的分布保持一致,因此大数据的统计特征可以用RSP数据块来估计,大数据的分类、聚类、回归等模型可以用随机抽取的少量RSP数据块来建立。Alpha计算框架提供了分布式环境下迭代渐近式的集成模型学习流程。任何大数据,一旦转换成RSP数据块后,大数据的分析就转换成RSP数据块的分析。因为单个RSP数据块就可以在集群的单个节点上独立计算,采用本文介绍的RSP数据模型和Alpha计算框架技术进行大数据分析极大地降低了集群计算资源的约束,提高了集群系统大数据处理与分析的扩展能力,在有限计算资源的集群上可以实现TB级大数据分析和建模能力。

对比现有的分布式大数据计算方法和分析技术,RSP数据模型在两个方面取得了突破:

(1)在分而治之的策略下,用随机抽取的少量RSP数据块计算取代了大数据所有数据块的计算,提高了计算效率和扩展能力;

(2)由于大数据的随机样本已经预先生成,不再需要分布式环境面向大数据所有记录的简单随机抽样操作。如果需要随机样本数据,随机抽取RSP数据块就可得到,计算量大大地降低。

第一个问题是目前大数据分析的主要技术瓶颈,Hadoop MapReduce和Spark采用的HDFS文件系统存储大数据,由于HDFS的数据块不是大数据的随机样本,因此要取得正确的结果,必须计算整个大数据,计算能力受到计算资源限制,特别是内存资源的约束。

除上述两个基本突破外,RSP数据模型和Alpha计算框架还带来了如下两个优势:

(1)由于每个数据块在单个计算节点独立计算,现有的串行算法都可以直接使用,不再需要并行算法,降低了算法并行化的成本。

(2)可以实现RSP数据块存储系统与大数据分析平台的分离。因为采用Alpha计算框架,在分批计算基模型时,可以从RSP数据块存储系统中随机抽取少量RSP数据块,下载到分析平台建模和集成,不需要在存储RSP大数据的平台上直接运算,可以很好地实现大数据的存储共享。

RSP数据模型和Alpha计算框架是为TB级以上大数据的分析设计开发的新技术,许多大数据分

析任务[24-25]都可以采用此技术完成。目前笔者团队所完成的工作还只是实现一些基本功能,很多理论和技术问题需要深入研究解决。但是,初期的成果已经展示出这一新技术的发展前景,为大数据分析提供了一个新的可选择方案,可以促进大数据的分析与应用技术的发展。

参考文献:

[1] 陈国良. 大数据聚类专题序言[J]. 深圳大学学报(理工版), 2019, 36(1): 1-3.
Chen Guoliang. Editorial of special issue on big data clustering[J]. Journal of Shenzhen University Science and Engineering, 2019, 36(1): 1-3.

[2] 黄晓云. 基于HDFS的云存储服务系统研究[D]. 大连: 大连海事大学, 2010.
Huang Xiaoyun. Cloud storage service system based on HDFS[D]. Dalian: Dalian Maritime University, 2010.

[3] 付东华. 基于HDFS的海量分布式文件系统的研究与优化[D]. 北京: 北京邮电大学, 2012.
Fu Donghua. Research and improvement of the massive distributed file system based on HDFS[D]. Beijing: Beijing University of Posts and Telecommunications, 2012.

[4] Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51(1): 107-113.

[5] Dean J, Ghemawat S. MapReduce: A flexible data processing tool[J]. Communications of the ACM, 2010, 53(1): 72-77.

[6] 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.

[7] 张滨. 基于MapReduce大数据并行处理的若干关键技术研究[D]. 上海: 东华大学, 2017.
Zhang Bin. Research on some key technologies of parallel processing for big data based on MapReduce[J]. Shanghai: Donghua University, 2017.

[8] 李志斌. 基于MapReduce并行处理框架的大数据处理系统的研究[D]. 吉林: 吉林大学, 2018.
Li Zhibin. Research on big data processing system based on MapReduce parallel processing framework[D]. Jilin: Jilin University, 2018.

[9] 王晨曦, 吕方, 崔慧敏, 等. 面向大数据处理的基于Spark的异质内存编程框架[J]. 计算机研究与发展, 2018, 55(2): 246-264.
Wang Chenxi, Lü Fang, Cui Huimin, et al. Heterogeneous memory programming framework based on Spark for big data processing[J]. Journal of Computer Research and Development, 2018, 55(2): 246-264.

[10] 宋泊东, 张立臣, 江其洲. 基于Spark的分布式大数据分析算法研究[J]. 计算机应用与软件, 2019, 36(1): 39-44.
Song Bodong, Zhang Lichen, Jiang Qizhou. Distributed big data analysis algorithm based on Spark[J]. Computer Applications and Software, 2019, 36(1): 39-44.

[11] 吴信东, 嵇圣硙. MapReduce与Spark用于大数据分析之比较[J]. 软件学报, 2018, 29(6): 1770-1791.
Wu Xindong, Ji Shengwei. Comparative study on MapReduce and Spark for big data analytics[J]. Journal of Software, 2018, 29(6): 1770-1791.

[12] Salloum S, He Y L, Huang Z X, et al. A random sample partition data model for big data analysis[EB/OL]. [2018-01-02](2019-03-01). https://arxiv.org/abs/1712.04146.

[13] Gretton A, Borgwardt K M, Rasch M J, et al. A kernel two-sample test[J]. Journal of Machine Learning Research, 2012, 13: 723-773.

[14] 魏丞昊,黄哲学, 何玉林, 等. 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.
Wei Chenghao, Huang Zhexue, He Yuling, et al. Statistical aware based big data system computing framework[J]. Journal of Shenzhen University Science and Engineering, 2018, 35(5): 441-443.

[15] Wei C H, Salloum S, Emara T, et al. A two-stage data processing algorithm to generate random sample partitions for big data analysis[C]//2018 International Conference on Cloud Computing. Cypru: Springer, 2018: 347-364.

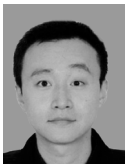
[16] Salloum S, Huang Z X, He Y L. Empirical analysis of asymptotic ensemble learning for big data[C]//2016 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies. Shanghai, China: IEEE, 2016: 8-17.

[17] Salloum S, Huang J Z, He Y L, et al. An asymptotic ensemble learning framework for big data analysis[J]. IEEE Access,

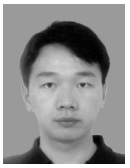
2019, 7: 3675-3693.
[18] Van der Vaart A W. Asymptotic statistics[M]. UK: Cambridge University Press, 2000.
[19] 蔡毅, 朱秀芳, 孙章丽, 等. 半监督集成学习综述[J]. 计算机科学, 2017, 44(S1): 7-13.
Cai Yi, Zhu Xiufang, Sun Zhangli, et al. Semi-supervised and ensemble learning: A review[J]. Computer Science, 2017, 44 (S1): 7-13.
[20] 杨草原, 刘大有, 杨博, 等. 聚类集成方法研究[J]. 计算机科学, 2011, 38(2): 166-170.
Yang Caoyuan, Liu Dayou, Yang Bo, et al. Research on cluster aggregation approaches[J]. Computer Science, 2011, 38(2): 166-170.
[21] Huang Z X, Ng M K. A note on k-modes clustering[J]. Journal of Classification, 2003, 20(2): 257-261.
[22] Huang Z X. Extensions to the k-means algorithm for clustering large data sets with categorical values[J]. Data Mining and Knowledge Discovery, 1998, 2(3): 283-304.
[23] 曹科研, 栾方军, 孙焕良, 等. 不确定数据基于密度的局部异常点检测[J]. 计算机学报, 2017, 40(10): 2231-2244.
Cao Keyan, Luan Fangjun, Sun Huanliang, et al. Survey on the management of uncertain data[J]. Chinese Journal of Computers, 2017, 40(10): 2231-2244.
[24] Vargas-Solar G, Zechinelli-Martini J L, Espinosa-Oviedo J A. Big data management: What to keep from the past to face future challenges[J]. Data Science and Engineering, 2017, 2(4): 328-345.
[25] Siuly S, Zhang Y. Medical big data: Neurological diseases diagnosis through medical data analysis[J]. Data Science and Engineering, 2016, 1(2): 54-64.

作者简介:

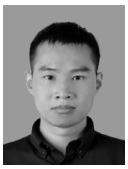
黄哲学(1959-),男,博士,教授,研究方向:大数据系统计算技术,E-mail: zx.huang@szu.edu. cn。



何玉林(1982-),男,博士,副研究员,研究方向:机器学习与数据挖掘,E-mail: yulinhe@szu.edu.cn。



魏丞昊(1986-),男,博士,副研究员,研究方向:Hadoop/Spark大数据处理,E-mail: chenghao.wei@szu.edu.cn。



张晓亮(1984-),男,博士,博士后,研究方向:大数据分析的数学理论与方法,E-mail:zhangxlas@163.com。

(编辑:张黄群)