DOI:10.3969/j.issn.1000-1565.2018.03.011

# 大数据与大数据机器学习

张素芳[1]，翟俊海[2]，王聪[2]，沈矗[2]，赵春玲[2]
(1.中国气象局气象干部培训学院河北分院，河北 保定 071000；
2.河北省机器学习与计算智能重点实验室，河北大学 数学与信息科学学院，河北 保定 071002)

**摘 要** 大数据时代已经到来，大数据是指具有海量(Volume)、多样(Variety)、时效(Velocity)、不精确(Veracity)和价值(Value)这 5 种特征的数据，大数据研究是近几年信息处理领域最热门的研究方向，已经引起了工业界、学术界乃至政府部门的高度关注.大数据之所以备受关注，是因为大数据里面蕴藏着巨大的价值.如何把蕴藏在大数据中的价值挖掘出来，为企业或政府部门提供决策支持具有重要的意义.大数据给传统的机器学习带来了许多挑战，这些挑战可以从大数据的 5 个特征或从 5 个不同的角度进行分析.本文首先介绍大数据的概念，并详细剖析大数据 5 种特征的内涵；然后在此基础上，重点分析大数据给机器学习带来的挑战及可能的解决方法.本文对从事大数据研究的人员，特别是从事大数据机器学习研究的人员具有较高的参考价值.

**关键词**：大数据；机器学习；云计算；决策支持
**中图分类号**：TP181 **文献标志码**：A **文章编号**：1000-1565(2018)03-0299-10

# Big data and big data machine learning

ZHANG Sufang[1], ZHAI Junhai[2], WANG Cong[2], SHEN Chu[2], ZHAO Chunling[2]
(1. Hebei Branch of China Meteorological Administration Training Centre, China Meteorological Administration, Baoding 071000, China;
2. Key Laboratory of Machine Learning and Computational Intelligence of Hebei Province, College of Mathematics and Information Science, Hebei University, Baoding 071002, China)

**Abstract**: Big data era has arrived. The big data refers to the data which is usually characterized by the 5 features: volume, variety, velocity, veracity, and value. In recent years, big data research is the hottest research topic in the field of information processing, and has drawn great attention from industrial communities, academic communities and governments because big value can be found in big data. It is of great significance for companies or governments to make decisions using the knowledge found from big data. Big data introduces many challenges to traditional machine learning, which can be analyzed by the 5 features of

**收稿日期**：2017-12-23
**基金项目**：国家自然科学基金资助项目(71371063)；河北省自然科学基金资助项目(F2017201026)；河北大学自然科学研究计划项目(799207217071)；河北大学研究生创新资助项目(hbu2018ss47)；河北大学大学生创新训练项目(2017071)
**第一作者**：张素芳(1966—)，女，河北蠡县人，中国气象局气象干部培训学院河北分院副教授，主要从事机器学习方向研究.E-mail: mczsf@126.com
**通信作者**：翟俊海(1964—)，男，河北易县人，河北大学教授，博士，主要从事机器学习和数据挖掘方向研究.E-mail:mczjh@126.com

big data or from 5 different views. This paper firstly introduces the concept of big data, and carefully analyzes the connotations of the 5 features, and then mainly focuses on analyzing the challenges and the possible solutions. This paper can be very helpful to researchers in related fields, especially for the ones engaging in the study of big data machine learning.

**Key words**: big data; machine learning; cloud computing; decision making

## 1 大数据及其 5Vs 特征

随着网络技术、数据存储技术和物联网技术的快速发展,以及移动通信设备的普及,数据正以前所未有的速度在增长,人类已经进入了大数据时代.那么究竟什么样的数据才是大数据呢?目前没有关于大数据的标准定义.狭义地讲,大数据就是海量数据,是指大小超过一定量级(如 TB 级、PB 级)的数据.美国著名的麦肯锡公司给出一个狭义的大数据定义[1]:大数据是指大小超出常规软件获取、存储、管理和分析能力的数据.狭义的定义只考虑了大数据的量级,没有考虑大数据的其他特征.广义地讲,大数据不只是量大的数据,还有其他的特征.目前,被广泛接受的是用 5 个特征定义的大数据[2-4],即大数据是指具有海量(Volume)、多样(Variety)、时效(Velocity)、不精确(Veracity)和价值(Value)这 5 种特征的数据,这 5 种特征简称大数据的 5Vs 特征.

在这 5 个特征中,价值特征处于核心位置,如图 1 所示.大数据之所以受到极大关注,就是因为大数据中蕴含着巨大的价值.

下面详细阐述大数据的这 5Vs 特征的含义.

海量性特征(Volume):是指数据量大,即所谓的海量.数据的量级已从 TB(1 TB=$2^{10}$ GB)级别转向 PB(1 PB=$2^{10}$ TB)量级,正在向 ZB(1 ZB=$2^{10}$ PB)量级转变.从机器学习的角度讲,量大有 2 种表现形式:1)是数据集中样例的个数超多;2)是表示样例的属性或特征的维数超高.

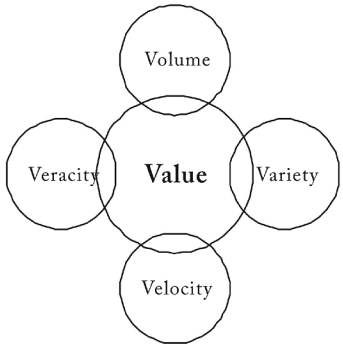


图 1 大数据的 5Vs 特征

Fig. 1 5Vs of big data

多样性特征(Variety):是指数据类型、表现形式和数据源多种多样.数据类型可能是结构化数据(如表结构的数据),也可能是无结构化数据(如文档数据),还可能是半结构化数据(如 Web 网页数据);数据的表现形式呈现出多种模态,如音频、视频、日志等.数据源可能是同构的,也可能是异构的.

时效性特征(Velocity):是指数据需要及时处理,否则数据就会失去其应用价值.随着网络技术、数据存储技术和物联网技术的快速发展,以及移动通信设备的普及,数据呈爆炸式快速增长,新数据不断涌现,快速增长的数据要求数据处理的速度也要快,这样才能使大量的数据得到有效的利用.在实践中,很多大数据都需要在一定时间内及时处理,例如电子商务大数据.

不精确性特征(Veracity):是指由数据的质量、可靠性、不确定性、不完备性引起的不确定性.这一特征有时也从其对立面考虑,称为数据的真实性.数据的重要性体现在其应用价值,数据的规模并不能决定其是否有应用价值,数据的真实性是保证能挖掘到具有应用价值或潜在应用价值的规律或规则的重要因素.

价值性特征(Value):是大数据的核心特征,它包括 2 层含义,一是指大数据的价值密度低,二是指大数据的确蕴含着巨大的价值.例如,用于罪犯跟踪的视频大数据,可能对罪犯跟踪有价值的只有很少的几个帧,但正是这关键的几帧数据,却有重大的价值.

## 2 机器学习及常用方法

机器学习[5-7]研究机器如何模拟或实现人类的学习行为,以获取新的知识或技能,重新组织已有的知识结构使之不断改善自身的性能.它是人工智能的核心,是使机器获取智能的根本途径.具体地,机器学习是通过各种算法(与学习任务相关)从数据中学习如何完成特定的任务,这个过程称为训练,然后用学到的知识(规律、规则或模型)对真实世界中的事件做出决策或预测.机器学习最近 20 年取得了巨大的进步,目前非常

火热的深度学习[8-9]是机器学习的一种特例，它可以看作是训练深度模型（如深度神经网络）的一种算法，深度学习在计算机视觉、自然语言处理、音频识别等领域获得了巨大的成功[10-11]．Google 公司开发的轰动全球的 3 个围棋程序 AlphaGo、Master 和 AlphaGo Zero 都是用深度学习算法训练的[12-13]．

机器学习可大致分为 3 类[14]：监督学习、无监督学习和强化学习，如图 2 所示．在监督学习中，训练样例带有类别标签（简称类标），可以表示成二元对$(x, y)$，其中，$y$ 是样例 $x$ 的类标．监督学习按着预定的学习准则，如要求均方误差最小或分类精度最高等，通过学习算法调整学习模型 $f(x)$ 中的参数，目的是学到最优模型 $f(x)$．然后用学到的模型 $f(x)$ 来预测新样例 $x$ 的类标 $y$（或在给定 $x$ 的情况下，输出 $y$ 的概率分布）．模型 $f(x)$ 有多种形式，包括神经网络、支持向量机、决策树、逻辑回归，贝叶斯分类器等．一般地，从训练集中学习模型 $f(x)$ 的过程都要用到优化方法或数值分析的方法．例如，在支持向量机中，要用到二次优化方法；在神经网络中，要用到梯度优化方法等．

无监督学习在学习的过程中，没有监督信息可以利用．无监督学习只处理“特征”，不操作监督信号，它通常与密度估计相关[15]．例如，学习从数据分布中采样、寻找数据分布的流形、将数据中相关的样本聚类等．无监督学习的任务是寻找数据的“最佳”表示，对于不同的问题，“最佳”表示的含义也不相同．例如，在主成分分析中，“最佳”表示的含义是寻找表示数据的最优投影子空间；而对于流形学习，“最佳”表示的含义是寻找接近数据真实分布的流形．无监督学习最常见的表现形式是聚类分析，它根据数据集自身的特性，将数据集中的样例划分为若干个簇，簇内的样例之间比不同簇的样例之间更相似．常见的聚类方法包括 $K$ 均值聚类、自组织映射、层次聚类等，如图 2 所示．在无监督学习中，针对给定的数据集，选择合适的相似性度量至关重要，常用的相似性度量包括基于距离的度量和基于相关性的度量[16]．

强化学习是指通过智能体（Agent）与环境的交互，以环境对智能体的反馈为输入，通过学习选择能达到其目标的最优动作，即学习的结果是一个最优策略．“试错搜索”和“延迟回报”是强化学习的 2 个主要的特征．大多数强化学习方法都是建立在马尔科夫决策过程（Markov Decision Process，MDP）理论框架之上的．根据智能体在学习过程中，是否需要学习 MDP 模型知识，强化学习算法可分为模型无关算法和模型相关算法，如图 2 所示．在通常情况下，模型无关算法每次迭代计算量小，且对动态环境适应性好，是强化学习中最主要的学习技术之一，其中，Q-学习算法是最常用的一个模型无关基本算法．

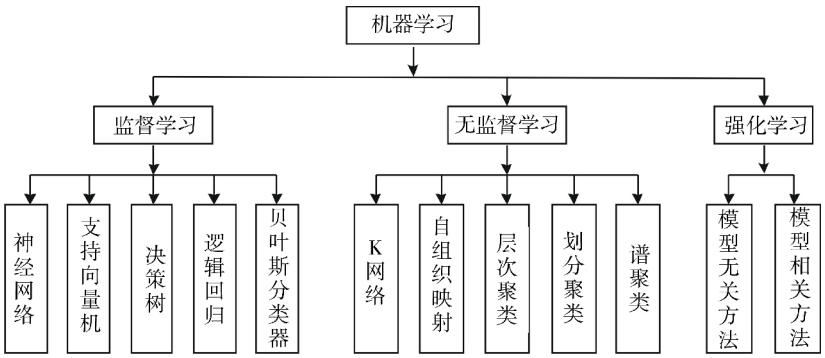


图 2　机器学习方法分类

Fig. 2　Categories of machine learning methods

## 3　大数据机器学习面临的挑战及可能的解决方法

虽然文献中已经有一些关于大数据机器学习综述的文献[22-26]，但不同于已有的文献，本文从全新的视角，即从大数据的 5Vs 特征这 5 种不同的视角，讨论大数据机器学习面临的挑战及可能的解决方法，包括笔者自己提出的一些方法．

### 3.1　海量性挑战

传统的机器学习算法都属于内存算法，用这些算法从训练集中学习时，需要将整个数据集加载到内存中．但是，在大数据时代，训练集的大小远远超过计算机的内存容量．在这种情况下，传统的机器学习算法变得不可行．针对这种挑战，有如下 2 种可能的解决方法．

### 3.1.1 基于并行计算的解决方法

这种解决方法的基本思想是分治策略,即把海量机器学习问题划分成一些中小型的机器学习问题,然后分而治之.需要注意的是,并不是所有的机器学习问题都能用并行化方法解决.如果一个大数据可以分为具有同样计算过程的数据块,并且这些数据块之间不存在数据依赖关系,则可以并行处理,相应的大数据机器学习问题就可以用并行化方法解决.文献[27]给出了一个判别准则:"满足'求和范式'的大数据机器学习问题都能用并行计算的方法解决",具有很高的参考价值.目前,主流的开源大数据处理平台,例如 Hadoop[28] 和 Spark[29] 都对大数据并行计算的解决方法有很好的支持.在 Hadoop 和 Spark 中,分别包含有针对机器学习的开源库 mahout[30] 和 MLlib[31].在这 2 个机器学习开源库中,集成了一些经典的机器学习算法.例如,逻辑回归、朴素贝叶斯分类器、决策树、随机森林、多层感知器、$K$-means、高斯混合模型、交替最小二乘推荐算法等.近几年,基于这 2 种开源平台的大数据机器学习研究,也是这类方法的研究热点.例如,最近非常火热的大数据深度学习[10,32-34],基于 MapReduce 的大数据关联规则挖掘[35],基于 MapReduce 的大数据聚类[36-38],基于 Spark 的协同过滤推荐方法[39],基于 Spark 的 $K$-近邻分类方法[40],基于 MapReduce 的大数据 $K$-近邻分类方法[41],基于 Hadoop 分布式支持向量机[42],基于 MapReduce 的大数据优化预测[43-44]等.

事实上,Hadoop 和 Spark 处理大数据的基本思想都是分治策略,而且都采用 HDFS(Hadoop Distributed File System)实现对大数据的分布式存储.Hadoop 采用 MapReduce 编程框架[45] 实现对大数据的处理.狭义地讲,Map 将大数据集划分为若干子集,并将这些子集部署到不同的云计算节点上,并行地对数据子集进行处理;Raduce 对各个云计算节点处理的中间结果进行进一步的处理,并得到最终结果.MapReduce 为用户提供了 2 个简单易行的编程接口:Map 和 Reduce,用户用它们去实现基本的并行计算.为了减少云计算节点之间的通信开销,MapReduce 还提供了 sort、combine、partition、merge 等辅助操作.以文本大数据词频统计为例,用 MapReduce 处理大数据的流程如图 3 所示.

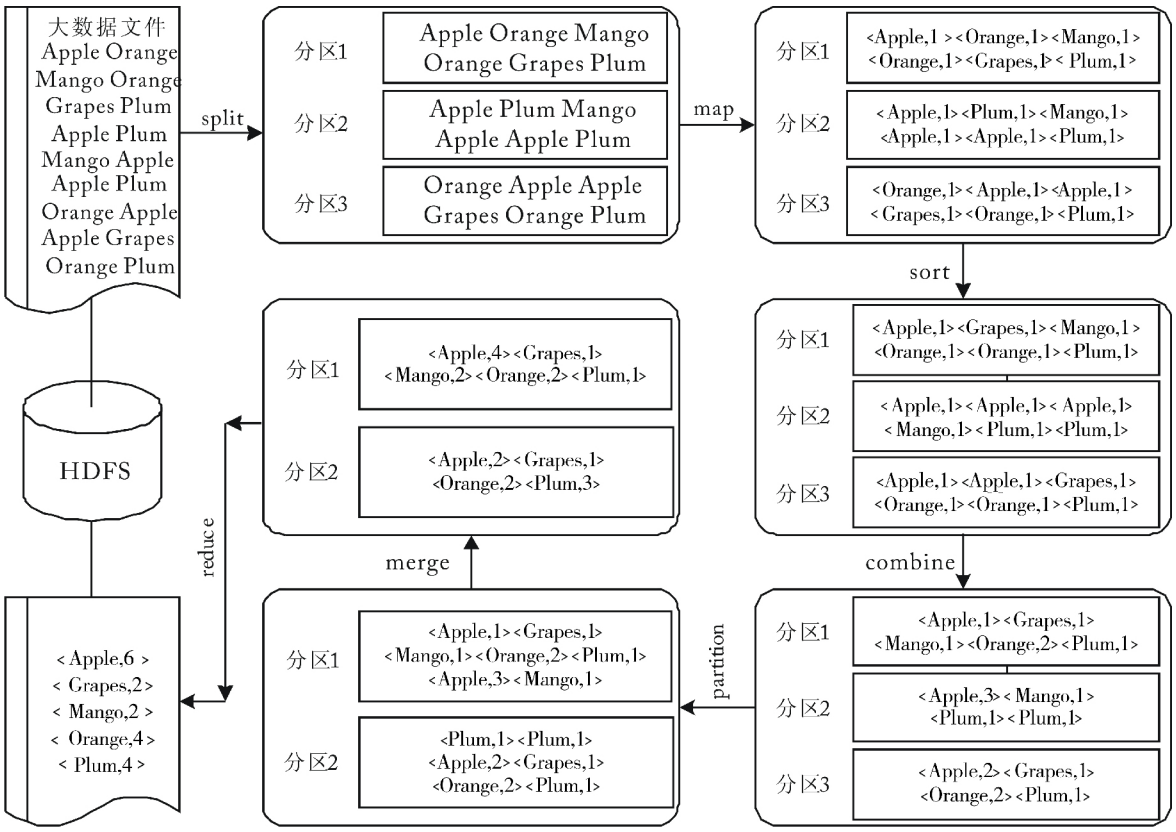


**图 3 用 MapReduce 处理大数据的流程**

**Fig.3 Big data processing procedure with MapReduce**

Spark 是在克服 Hadoop 不足的基础上提出的，它的首要设计目标是避免运算时出现过多的网络和磁盘 I/O 开销，为此它将核心数据结构设计为弹性分布式数据集 RDD(Resident Distributed Dataset)，Spark 使用 RDD 实现基于内存的计算框架，在计算过程中它会优先考虑将数据缓存在内存中，如果内存容量不足的话，Spark 才会考虑将数据缓存到磁盘上或者部分数据缓存到磁盘上．Spark 为 RDD 提供了一系列算子，以对 RDD 进行有效的操作．此外，为了避免 Hadoop 启动和调度作业消耗过大的问题，Spark 采用基于有向无环图 DAG(Directed Acyclic Graph)的任务调度机制进行优化，这样可以将多个阶段的任务并行或串行执行，无需将每一个阶段的中间结果存储到 HDFS(Hadoop Distributed File System)上．仍以文本大数据词频统计为例，用 Spark 处理大数据的流程如图 4 所示．从图 3 和图 4 可以看出，Hadoop 和 Spark 处理大数据的流程非常相似．

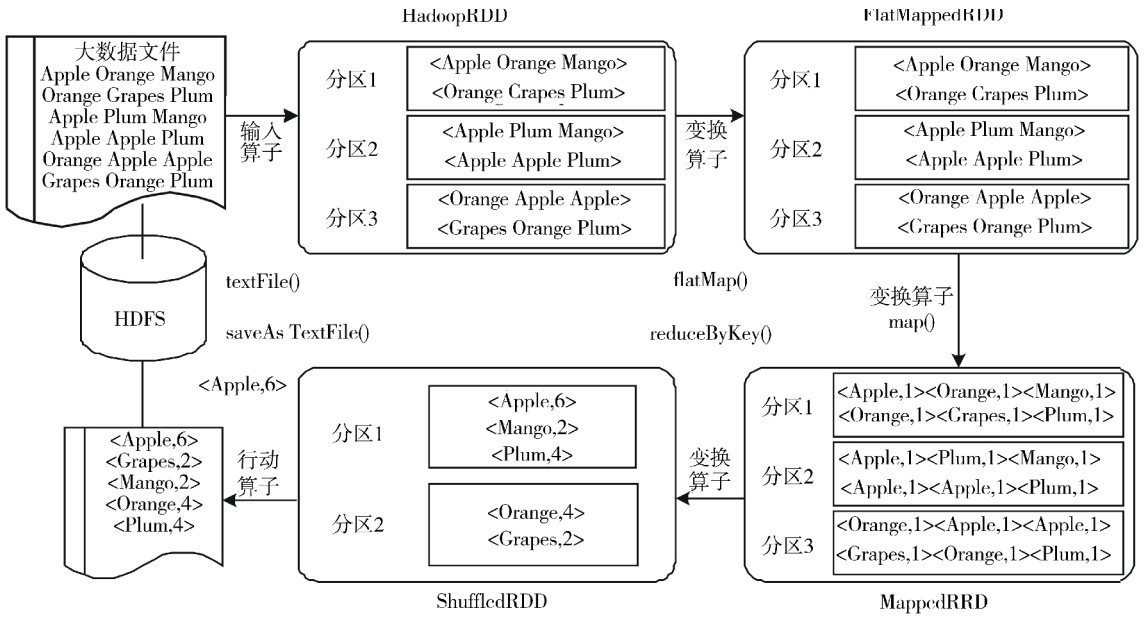


**图 4　Spark 大数据处理流程**

**Fig. 4　Big data processing procedure with Spark**

### 3.1.2　基于数据约简的解决方法

这类方法也称为样例选择方法，是从大数据集中选择一个有代表性的子集，代替大数据集进行机器学习．针对中小型数据集的样例选择算法已有很多，但针对大数据集的样例选择算法相对较少[46]．针对大数据分类问题，Triguero 等[47]提出了一种原型约简的 MapReduce 解决方案．基于局部敏感哈希技术，Alvar 等[48]提出了计算时间复杂度为线性级的样例选择算法，这种方法能实现大数据集的样例选择．基于随机突变爬山算法，Si 等[49]提出了一种用 MapReduce 进行大数据样例选择的算法．基于 MapReduce 和投票机制，Thai 等提出了一种大数据样例选择方法[50]．这种方法的基本思想如下：首先利用 MapReduce 的 Map 机制，将大数据集划分为若干个子集，并部署到不同的云计算节点上．然后在这些云计算节点上，并行地从子集上选择样例．接下来利用 MapReduce 的 Reduce 机制，合并这些云计算节点选择的样例子集，得到一次选择的样例子集．重复上述过程若干次，得到若干个选择的样例子集，最后投票选出最重要的样例子集．这种方法的优点是不依赖于任何样例选择算法，具有通用性．图 5 给出了一个示意性的例子，说明了用该算法进行大数据样例选择的过程．

## 3.2　多样性挑战

多样性的挑战主要体现在多模态大数据机器学习上，数据呈现为视频、音频、文本等多种模态，针对这种挑战可能的解决方法是数据融合或集成学习．因为不同模态的数据之间既相互区别，又相互联系，所以如何有效地融合不同模态的大数据，以实现高效的多模态大数据学习是解决这种挑战的关键．目前，主流的融合技术有 2 种：第 1 种是直接对不同模态的数据在原始空间进行融合；第 2 种是对不同模态的数据进行哈希变

换,然后在变换空间进行融合.在第 1 种方法中,比较有代表性的研究工作包括:基于深度神经网络的多模态大数据融合[51-52]和基于语义的多模态大数据融合[53],文献[53]对多模态大数据融合的研究方法进行了全面而深入的综述,具有很高的参考价值.第 2 种方法的基本思想是对原始的数据集进行哈希变换,将其变换到哈希空间(也称为海明空间).因为在海明空间,每一个对象都用一个 0-1 串表示,所以可以极大地降低数据复杂度.针对多模态大数据的哈希学习,最主要的问题是在构建哈希函数时,如何有效利用来自多个模态的信息.Zhang 等[54]提出了一种处理多源数据的加权复合哈希学习方法,通过对各个来源的数据赋予不同的权重来获得最终统一的哈希编码.Wu 等[55]研究了集成图像特征和文本特征的多模态哈希学习问题,虽然他们对图像特征和文本特征分别构建哈希函数,但哈希函数的学习通过一个共同的目标函数完成,从而实现了交叉模态样例的索引.Liu 等[56]提出了一种基于 Boosting 技术的多模态哈希学习方法,该方法通过迭代的方式产生一系列具有互补特征的多个哈希表,从而获得了较好的检索准确率,但是需要存储多个哈希表,增加了额外的存储开销.针对多模态大数据分类问题,本文也提出了自己的解决思路,即基于模糊积分的成对模态深度学习集成方法.该方法的技术路线如图 6 所示.

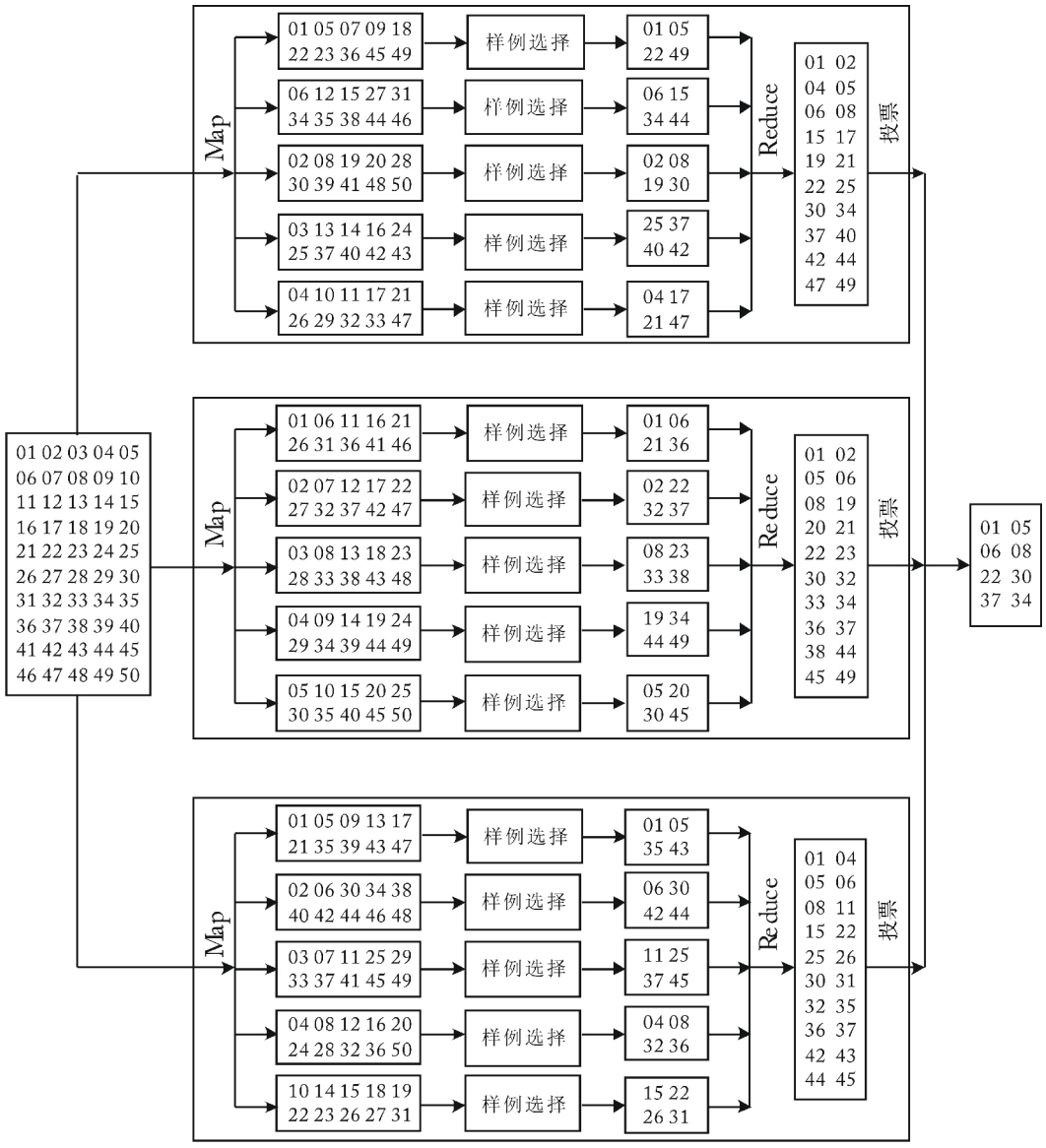


**图 5 基于 MapReduce 和投票策略的大数据样例选择过程**

**Fig.5 Schematic diagram of big data instance selection based on MapReduce and voting strategy**

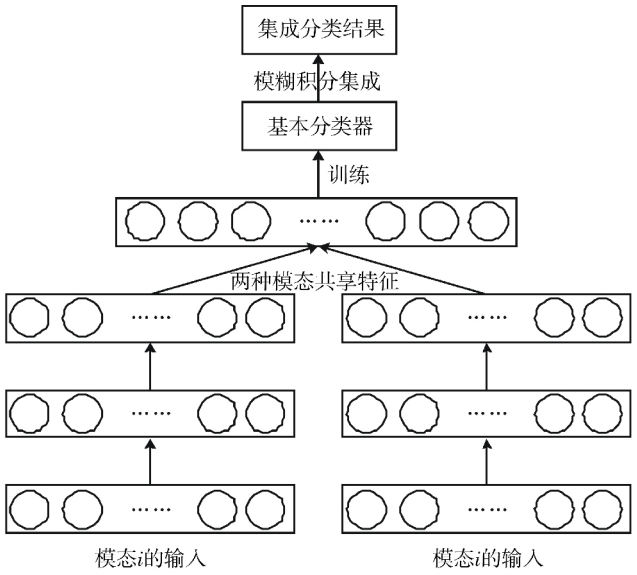


**图 6　多模态大数据集成分类**

**Fig.6　Multi-modal big data classification based on ensemble learning**

### 3.3　时效性挑战

时效性挑战是指大数据需要在一定时间内及时处理,否则数据就会失去其应用价值.解决这种挑战可能的方法是在线学习(Online Learning)和流式学习(Streaming Learning).正是由于有这种需求,开源大数据处理平台 Spark 提供了流学习库(Spark Streaming)[29].目前,文献中出现的解决这类挑战的方法大都是基于 Spark 的流学习库.例如,Ramirez-Gallego 等[57]利用 Spark 的流式学习库,提出了针对高速大数据流的最近邻分类方法.Lekha 等[58]提出了基于 Spark 的流式大数据机器学习方法,并应用于人类健康状态预测,解决了时效性问题.Carcilloa 等[59]利用 Spark 提出了一种针对流式大数据学习的可扩展学习框架,并应用于信用卡欺诈检测.此外,Wu 等[60]开发了一个称为 SOL(Scalable Online Learning)的大数据在线学习库.Cong 等[61]利用过拟合机制提出了一种在线大数据相似性学习方法.基于在线序列极限学习机[62],Zhai 等[63]提出了一种针对大规模数据集分类的模糊积分集成方法.

### 3.4　不精确性挑战

不精确性挑战是指大数据的质量、可靠性、不确定性、不完备性引起的不确定性挑战.解决这种挑战可能的方法是基于不确定性理论(如模糊集理论)的方法.但是,目前这方面的文献还非常少,有兴趣的读者可参考文献[64]和[65],在这 2 篇文献中对大数据中的不确定性及其作用进行了讨论,具有较高的参考价值.

### 3.5　价值性挑战

价值性挑战是指大数据的价值密度低,挖掘非常困难的挑战.解决这种挑战可能的方法是非平衡大数据学习方法.针对这种挑战的研究,即针对非平衡大数据学习研究还相对较少,只有较少的一些研究人员进行了这方面研究.Rio 等[66]利用随机森林作为分类器,对传统的处理类别非平衡的方法(随机上采样、下采样和代价敏感性学习)进行并行化改进,使这 3 种方法可以适用于非平衡大数据环境.Ghanavati 等[67]对传统的 SMOTE 方法进行改进,提出了一种组合度量学习和平衡化技术的非平衡大数据学习方法.D'Addabbo 等[68]提出了一种并行选择性采样方法,该方法以支持向量机(SVM: Support Vector Machine)作为分类器,利用 Tomek 链接作为启发式从负类样例中进行下采样,以实现非平衡大数据的分类.在 MapReduce 框架下,Lopez 等[69]提出了一种针对非平衡大数据的代价敏感性语义模糊规则分类算法,该算法利用 MapReduce 分布式计算代价敏感性模糊分类模型.基于 MapReduce 和集成学习,Zhai 等[70]提出了一种非平衡大数据分类方法.Fernandez 等[71]对非平衡大数据分类问题进行了简单的综述,具有一定的参考价值.

## 4 结论

首先介绍了大数据的概念,并详细剖析大数据 5 种特征的内涵;然后从大数据的 5 Vs 特征这 5 种全新的视角,分析了大数据机器学习面临的挑战及可能的解决方案,也包括作者所在研究团队提出的解决方案.在这 5 种挑战中,研究人员对海量性挑战的研究最多,成果也最为丰富.针对第 2 种和第 3 种挑战的研究次之,已有一些研究成果,但不是太多.针对另外 2 种挑战的研究,还相对较少.但是,作者发现针对非平衡大数据学习的研究已经引起了研究人员的关注,将会成为未来几年大数据机器学习的一个研究热点.这主要因为很多现实大数据处理问题本身就是非平衡大数据学习问题.例如,医疗大数据处理问题、罪犯视频跟踪大数据处理问题、信用卡欺诈检测大数据处理问题等.

**参 考 文 献:**


[1] MANYIKA J, CHUI M, BROWN B, et al. Big data: The next frontier for innovation, competition, and productivity [R/OL]. https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/ big-data-the-next-frontier-for-innovation.

[2] EMANI C K, CULLOT N, NICOLLE C. Understandable Big Data: A survey [J]. Computer Science Review, 2015, 17: 70-81. DOI: 10.1016/j.cosrev.2015.05.002.

[3] **孟小峰, 慈祥. 大数据管理:概念、技术与挑战** [J]. **计算机研究与发展**, 2013, 50(1):146-169. DOI:10.7544/issn1000-1239.2013.20121130.
MENG X F, CI X. Big data management: concept, techniques and challenges [J]. Journal of Computer Research and Development, 2013, 50(1):146-169. DOI:10.7544/issn1000-1239.2013.20121130.

[4] STOREY V C, SONG I Y. Big data technologies and management: What conceptual modeling can do [J]. Data & Knowledge Engineering, 2017, 108:50-67. DOI: 10.1016/j.datak.2017.01.001.

[5] MITCHELL T M. **机器学习**[M].**英文影印版.北京: 机械工业出版社**, 2003.

[6] MURPHY K. Machine learning: a probabilistic perspective [M]. Cambridge: MIT Press, 2012.

[7] **周志华. 机器学习**[M]. **北京: 清华大学出版社**, 2016.

[8] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313:504-507. doi:10.1126/science.1127647.

[9] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553):436-444. DOI:10.1038/nature14539.

[10] **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. DOI: 10.11992/tis.201611021.
MA S L, WUNIRI Q Q G, LI X P. Deep learning with big data: state of the art and development [J]. CAAI Transactions on Intelligent Systems, 2016, 11(6):728-742. DOI: 10.11992/tis.201611021.

[11] GUO Y M, LIU Y, OERLEMANS A, et al. Deep learning for visual understanding: a review [J]. Neurocomputing, 2016, 187:27-48. DOI: 10.1016/j.neucom.2015.09.116.

[12] 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. DOI: 10.1038/nature16961.

[13] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of Go without human knowledge [J]. Nature, 2017, 550(7676):354-359. DOI: 10.1038/nature24270.

[14] JORDAN M I, MITCHELL T M. Machine learning: Trends, perspectives, and prospects [J]. Science, 2015, 349(6245):255-260. DOI: 10.1126/science.aaa8415.

[15] **赵申剑, 黎彧君, 符天凡, 等. 深度学习** [M].**北京: 人民邮电出版社**, 2017.

[16] CHA S H. Comprehensive survey on distance/similarity measures between probability density functions [J]. International Journal of Mathematical Models and Methods in Applied Sciences, 2007, 4(1):300-307.

[17] **董西成**. Hadoop **技术内幕** [M]. **北京: 机械工业出版社**, 2013.

[18] 黄宜华，苗凯翔. 深入理解大数据:大数据处理与编程实践[M]. 北京：机械工业出版社，2014.

[19] 刘军，林文辉，方澄. Spark 大数据处理－原理、算法与实例[M]. 清华大学出版社，2016.

[20] 樊哲. Mahout 算法解析与案例实战[M]. 北京：机械工业出版社，2014.

[21] NICK P. Spark 机器学习[M].影印版.北京：人民邮电出版社，2015.

[22] 何清，李宁，罗文娟,等. 大数据下的机器学习算法综述[J]. 模式识别与人工智能，2014，27(4):327－336. DOI:10.3969/j.issn.1003－6059.2014.04.007.
HE Q，LI N，LUO W J，et al. A survey of machine learning algorithms for big data [J]. Pattern Recognition and Artificial Intelligence，2014，27(4):327－336. DOI:10.3969/j.issn.1003－6059.2014.04.007.

[23] 黄宜华. 大数据机器学习系统研究进展[J]. 大数据，2015，1(1):28－47. DOI:10.11959/j.issn.2096－0271.2015004.
HUANG Y H. Research progress on big data machine learning system [J]. Big Data，2015，1(1):28－47. DOI:10.11959/j.issn.2096－0271.2015004.

[24] HEUREUX A，GROLINGER K，ELYAMANY H F，et al. Machine learning with big data：Challenges and approaches [J]. IEEE Access，2017，5:7776－7797. DOI：10.1109/ACCESS.2017.2696365.

[25] ZHOU L，PAN S，WANG J，et al. Machine learning on big data：Opportunities and challenges [J]. Neurocomputing，2017，237:350－361. DOI:10.1016/j.neucom.2017.01.026.

[26] AL－JARRAH O Y，YOO P D，MUHAIDAT S，et al. Efficient machine learning for big data：a review [J]. Big Data Research，2015，2(3):87－93. DOI:10.1016/j.bdr.2015.04.001.

[27] CHU C T，SANG K K，LIN Y A，et al. Map－reduce for machine learning on multicore [Z]. International Conference on Neural Information Processing Systems，Vancouver，Canada，2006.

[28] Apache. Hadoop [Z/OL]. [2017－12－01]. http://hadoop.apache.org/.

[29] Apache. Spark [Z/OL]. [2017－12－05]. http://spark.apache.org/.

[30] Apache. Mahout [Z/OL]. [2017－12－07]. http://mahout.apache.org/.

[31] Apache. MLlib [Z/OL]. [2017－12－12]. http://spark.apache.org/mllib/.

[32] CHEN X W，LIN X. Big data deep learning：challenges and perspectives [J]. IEEE Access，2014，2:514－525. DOI：10.1109/ACCESS.2014.2325029.

[33] ZHANG K，CHEN X W. Large－scale deep belief nets with MapReduce [J]. Access IEEE，2015，2(2):395－403. DOI：10.1109/ACCESS.2014.2319813.

[34] LV Y，DUAN Y，KANG W，et al. Traffic flow prediction with big data：a deep learning approach [J]. IEEE Transactions on Intelligent Transportation Systems，2015，16(2):865－873. DOI：10.1109/TITS.2014.2345663.

[35] BECHINI A，MARCELLONI F，SEGATORI A. A MapReduce solution for associative classification of big data [J]. Information Sciences，2016，332:33－55. DOI:10.1016/j.ins.2015.10.041.

[36] LUDWIG S A. MapReduce－based fuzzy c－means clustering algorithm：implementation and scalability [J]. International Journal of Machine Learning & Cybernetics，2015，6(6):923－934. DOI:10.1007/s13042－015－0367－0.

[37] LI X，SONG J，ZHANG F，et al. MapReduce－based fast fuzzy C－means algorithm for large－scale underwater image segmentation [J]. Future Generation Computer Systems，2016，65:90－101. DOI:10.1016/j.future.2016.03.004.

[38] XU Y，QU W，LI Z，et al. Efficient $K$－means＋＋ approximation with MapReduce [J]. IEEE Transactions on Parallel & Distributed Systems，2014，25(12):3135－3144. DOI:10.1109/TPDS.2014.2306193.

[39] PANIGRAHI S，LENKA R K，STITIPRAGYAN A. A hybrid distributed collaborative filtering recommender engine using Apache Spark [J]. Procedia Computer Science，2016，83:1000－1006. DOI:10.1016/j.procs.2016.04.214.

[40] MAILLO J，RAMíREZ S，TRIGUERO I，et al. kNN－IS：an iterative spark－based design of the $K$－nearest neighbors classifier for big data [J]. Knowledge－Based Systems，2017，117:3－15. DOI:10.1016/j.knosys.2016.06.012.

[41] 翟俊海，王婷婷，张明阳,等. 2 种加速 $K$－近邻方法的实验比较[J]. 河北大学学报(自然科学版)，2016，36(6):650－656. DOI:10.3969/j.issn.1000－1565.2016.06.013.
ZHAI J H，WANG T T，ZHANG M Y，et al. Experimental comparison of two acceleration approaches for $K$－nearest neighbors [J]. Journal of Hebei University (Natural Science Edition)，2016，36(6):650－656. DOI:10.3969/j.issn.1000－1565.2016.06.013.

[42] 高学伟，付忠广，孙力，等. 基于 Hadoop 分布式支持向量机球磨机大数据建模[J]. 河北大学学报(自然科学版)，2017，37(3):309－315. DOI:10.3969/j.issn.1000－1565.2017.03.014.

GAO X W, FU Z G, SUN L, et al. Big data modeling of ball mill based on distributed support vector machine on Hadoop platform [J]. Journal of Hebei University (Natural Science Edition), 2017, 37(3):309－315. DOI:10.3969/j.issn.1000－1565.2017.03.014.

[43] 罗文劼, 袁方, 杨秀丹. 基于建模技术构建运用大数据分析优化政务的环境[J]. 河北大学学报(自然科学版), 2017, 37(1):101－107. DOI:10.3969/j.issn.1000－1565.2017.01.015.
LUO W J, YUAN F, YANG X D. Building platform for optimizing E－government business using big data analysis based on modeling technique [J]. Journal of Hebei University (Natural Science Edition), 2017, 37(1):101－107. DOI:10.3969/j.issn.1000－1565.2017.01.015.

[44] 马国富, 王子贤, 马胜利. 基于大数据的服刑人员危险性预测[J]. 河北大学学报(自然科学版), 2016, 36(6):657－666. DOI:10.3969/j.issn.1000－1565.2016.06.014.
MA G F, WANG Z X, MA S L. Prediction of the risk of offenders based on big data [J]. Journal of Hebei University (Natural Science Edition), 2016, 36(6):657－666. DOI:10.3969/j.issn.1000－1565.2016.06.014.

[45] DEAN J, GHEMAWAT S. MapReduce: simplified data processing on large clusters [J]. Communications of the ACM, 2008, 51(1):107－113. DOI: 10.1145/1327452.1327492.

[46] OLVERA－LÓPEZ J A, CARRASCO－OCHOA J A, MARTiNEZ－TRINIDAD J F, et al. A review of instance selection methods [J]. Artificial Intelligence Review, 2010, 34(2):133－143. DOI:10.1007/s10462－010－9165－y.

[47] TRIGUERO I, PERALTA D, BACARDIT J, et al. MRPR: A MapReduce solution for prototype reduction in big data classification [J]. Neurocomputing, 2015, 150:331－345. DOI:10.1016/j.neucom.2014.04.078.

[48] ALVAR A G, JOSE－FRANCISCO D P, RODRiGUEZ J J, et al. Instance selection of linear complexity for big data [J]. Knowledge－Based Systems, 2016, 107:83－95. DOI: 10.1016/j.knosys.2016.05.056.

[49] SI L, YU J, WU W, et al. RMHC－MR: Instance selection by random mutation hill climbing algorithm with MapReduce in big data [J]. Procedia Computer Science, 2017, 111:252－259. DOI: 10.1016/j.procs.2017.06.061.

[50] ZHAI J H, WANG X Z, PANG X H. Voting－based instance selection from large data sets with Mapreduce and random weight networks [J]. Information Sciences, 2016, 367: 1066－1077. DOI:10.1016/j.ins.2016.07.026.

[51] SRIVASTAVA N, SALAKHUTDINOV R. Multimodal learning with deep boltzmann machines [J]. Journal of Machine Learning Research, 2014, 15(8):1967－2006.

[52] NGIAM J, KHOSLA A, KIM M, et al. Multimodal deep learning [Z].International Conference on Machine Learning, Washington, USA, 2011.

[53] ZHENG Y. Methodologies for cross－domain data fusion: an overview [J]. IEEE Transactions on Big Data, 2015, 1(1):16－34. DOI: 10.1109/TBDATA.2015.2465959.

[54] ZHANG D, WANG F, SI L. Composite hashing with multiple information sources [Z]. The 34th international ACM SIGIR conference on Research and development in Information Retrieval, Beijing, China, 2011. DOI: 10.1145/2009916.2009950.

[55] WU B T, YANG Q, ZHENG W S, et al. Quantized correlation hashing for fast cross－modal search [Z]. International Joint Conferences on Artificial Intelligence, Buenos Aires, Argentina, 2015.

[56] LIU X, HUANG L, DENG C, et al. Multi－view complementary hash tables for nearest neighbor search [Z]. IEEE International Conference on Computer Vision, Santiago Chile, 2015. DOI: 10.1109/ICCV.2015.132.

[57] RAMIREZ－GALLEGO S, KRAWCZYK B, GARCIA S, et al. Nearest neighbor classification for high－speed big data streams using spark [J]. IEEE Transactions on Systems Man & Cybernetics Systems, 2017, 47(10):2727－2739. DOI: 10.1109/TSMC.2017.2700889.

[58] LEKHA R N, SUJALA D S, SIDDHANTH D S. Applying spark based machine learning model on streaming big data for health status prediction [J]. Computers & Electrical Engineering. DOI: 10.1016/j.compeleceng.2017.03.009.

[59] CARCILLOA F, POZZOLOA A D, BORGNEA Y A L, et al. SCARFF : A scalable framework for streaming credit card fraud detection with spark [J]. Information Fusion, 2018, 41:182－194. DOI: 10.1016/j.inffus.2017.09.005.

[60] WU Y, HOI S C H, LIU C, et al. SOL: A library for scalable online learning algorithms [J]. Neurocomputing, 2017, 260:9－12. DOI: 10.1016/j.neucom.2017.03.077.

[61] CONG Y, LIU J, FAN B, et al. Online similarity learning for big data with overfitting [J]. IEEE Transactions on Big Data, 2017. DOI: 10.1109/TBDATA.2017.2688360.

(下转第336页)

[13] GRANOVETTER M. Economic action and social structure: the problem of embeddedness [J]. American Journal of Sociology,1985,91(3):481－510.

[14] 齐晓云.信息技术融合及其对组织绩效影响的实证研究[D].长春:吉林大学,2011.

[15] HUBER G P. Organizational learning: the contributing processes and the literatures [J].Organization Science,1991,(2):88－115.

[16] LARSON A. Network dyads in entrepreneurial settings: a study of the governance of exchange relationships [J]. Administrative Science Quarterly,1992,37:76－144.

[17] 孙国强,闫慧丽.网络组织治理机制对治理能力影响的实证研究[J].高等财经教育研究,2015,18:31－49.
SUN G Q,YAN H L. Research on impact of the network governance mechanisms on governance capability[J]. Research on Higher Education of Finance and Economics,2015,18:31－49.

[18] NAMBISAN S. Information technology and product service innovation: A brief assessment and some suggestions for future research [J]. Journal of the Association for Information System,2013 ,14 (4) :215－226.

[19] PAMELA J S. Business in the cloud: Research questions on governance, audit, and assurance [J]. Journal of Information Systems,2016,30(3):173－189.

(责任编辑:孟素兰)

(上接第 308 页)

[62] LIANG N Y, HUANG G B, SARATCHANDRAN P, et al. A fast and accurate online sequential learning algorithm for feedforward networks [J]. IEEE Transactions on Neural Networks, 2006, 17(6):1411－23. DOI: 10.1109/TNN.2006.880583.

[63] ZHAI J H, WANG J G, HU W X. Combination of OSELM classifiers with fuzzy integral for large scale classification [J]. Journal of Intelligent & Fuzzy Systems, 2015, 28(5):2257－2268. DOI: 10.3233/IFS－141508.

[64] WANG H, XU Z S, PEDRYCZ W. An overview on the roles of fuzzy set techniques in big data processing: Trends, challenges and opportunities [J]. Knowledge－Based Systems, 2017, 118:15－30. DOI: 10.1016/j.knosys.2016.11.008.

[65] MAUGIS F A G. Big data uncertainties [J]. Journal of Forensic and Legal Medicine, 2016. DOI: 10.1016/j.jflm.2016.09.005.

[66] HERRERA F. On the use of MapReduce for imbalanced big data using Random Forest [J]. Information Sciences, 2014, 285:112－137. DOI: 10.1016/j.ins.2014.03.043.

[67] GHANAVATI M, WONG R K, CHEN F, et al. An effective integrated method for learning big imbalanced data [Z]. IEEE International Congress on Big Data, Alaska, USA, 2014. DOI: 10.1109/BigData.Congress.2014.102.

[68] DADDABBO A, MAGLIETTA R. Parallel selective sampling method for imbalanced and large data classification [J]. Pattern Recognition Letters, 2015, 97:61－67. DOI: 10.1016/j.patrec.2015.05.008.

[69] LOPEZ V, DEL RIO S, BENITEZ J M, et al. Cost－sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data [J]. Fuzzy Sets and Systems, 2015, 258:5－38. DOI: 10.1016/j.fss.2014.01.015.

[70] ZHAI J H, ZHANG S F, WANG C X. The classification of imbalanced large data sets based on MapReduce and ensemble of ELM classifiers [J]. International Journal of Machine Learning & Cybernetics, 2017, 8(3):1009－1017. DOI: 10.1007/s13042－015－0478－7.

[71] FERNANDEZ A, RIO S D, CHAWLA N V, et al. An insight into imbalanced big data classification: outcomes and challenges [J]. Complex & Intelligent Systems, 2017, 3(2):105－120. DOI: 10.1007/s40747－017－0037－9.

(责任编辑:孟素兰)