大数据下的机器学习算法综述*

何 清[1] 李 宁[1,2,3] 罗文娟[1,2] 史忠植[1]

[1](中国科学院计算技术研究所 智能信息处理重点实验室 北京 100190)
[2](中国科学院大学 北京 100049)
[3](河北大学 数学与计算机学院 保定 071002)

摘 要 随着产业界数据量的爆炸式增长,大数据概念受到越来越多的关注.由于大数据的海量、复杂多样、变化快的特性,对于大数据环境下的应用问题,传统的在小数据上的机器学习算法很多已不再适用.因此,研究大数据环境下的机器学习算法成为学术界和产业界共同关注的话题.文中主要分析和总结当前用于处理大数据的机器学习算法的研究现状.此外,并行是处理大数据的主流方法,因此介绍一些并行算法,并引出大数据环境下机器学习研究所面临的问题.最后指出大数据机器学习的研究趋势.

关键词 大数据,机器学习,分类,聚类,并行算法
中图法分类号 TP 391

A Survey of Machine Learning Algorithms for Big Data

HE Qing[1], LI Ning[1,2,3], LUO Wen-Juan[1,2], SHI Zhong-Zhi[1]
[1](*Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)
[2](*University of Chinese Academy of Sciences, Beijing* 100049)
[3](*College of Mathematics and Computer Science, Hebei University, Baoding* 071002)

ABSTRACT

With the explosive growth of the industry data, more and more attention is paid to big data. However, due to the volume, complex and fast-changing characteristics of big data, traditional machine learning algorithms for small data are not applicable. Therefore, developing machine learning algorithms for big data is a research focus. In this paper, the state-of-the-art machine learning techniques for big data are introduced and analyzed. As parallelism is a mainstream strategy for applying machine learning algorithms to big data, some parallelism strategies are described in detail as well. Finally, the challenges of applying machine learning to big data and some interesting research trends of machine learning in big data are pointed out.

Key Words Big Data, Machine Learning, Classification, Clustering, Parallel Algorithm

* 国家自然科学基金项目(No. 61175052,61203297,61035003,61363058)、国家 863 计划项目(No. 2014AA012205,2013AA01A606,2012AA011003)资助
收稿日期:2013 - 06 - 05
作者简介 何清,男,1965 年生,研究员,博士生导师,主要研究方向为机器学习、数据挖掘、基于云计算的海量数据挖掘.李宁(通讯作者),女,1982 年生,博士研究生,主要研究方向为文本挖掘、机器学习.E-mail:lin@ics.ict.ac.cn.罗文娟,女,1987 年生,博士研究生,主要研究方向为文本挖掘、机器学习.史忠植,男,1941 年生,研究员,博士生导师,主要研究方向为人工智能.

1 引 言

随着产业界数据量的爆炸式增长，数据以前所未有的速度积累，大数据(Big Data)概念受到越来越多的关注. 大数据正在给数据密集型企业带来丰厚的利润，据估计仅 Google 公司在 2009 年就为美国经济贡献 540 亿美元[1]. 学术界和产业界关于大数据的认识也在逐步清晰化并形成共识.

2011 年的语义技术信息(Semantic Technologies Information，STI)峰会上，与会的语义网络和数据库学者讨论了大数据时代语义网络、语义技术及数据库领域所面临的挑战. 关于大数据带来的挑战主要有如下观点.

Brodie[2] 认为在真实、无模式和复杂的大数据或大数据语义网络中进行有意义的数据集成需多学科多技术交叉. Bizer 列举大数据时代的 Web 数据研究的 3 种挑战：1) Web 数据的拓扑结构，因为互联网中有大量的不同形式的数据存在，各种类型的数据都非常巨大；2) Web 数据的特点，这对于数据集成和大数据处理来说是一个值得研究的问题；3) 已有一些公开的可用的预先爬取好的 Web 数据集可用于评测和实验. 他认为未来的挑战都将围绕数据集成、大规模资源描述框架(Resource Description Framework，RDF)处理和数据质量评定. Boncz 认为如果人们想更广泛地使用语义网络，存在两大挑战：1)缺乏好的使用案例；2)现有的数据集成方法使得创建链接非常困难. Erling 认为大数据时代语义的价值体现在让数据集成驱动数据库管理系统(Database Management System，DBMS)的技术.

产业方面，大数据是现有产业升级与新产业诞生的重要推动力量. 大数据时代的到来，产业界需求与关注点发生重大转变：企业关注的重点转向数据，计算机行业正在转变为真正的信息行业，从追求计算速度转变为关注大数据处理能力，软件也将从编程为主转变为以数据处理为主. 大数据处理的兴起也改变云计算的发展方向，使其进入以分析即服务(Analytics as a Service，AaaS)为主要标志的 Cloud 2.0 时代[3].

机器学习算法在学术界和产业界都有巨大的实用价值. 由于大数据的大量、复杂特性，对于大数据下的应用问题，传统的在小数据上的机器学习算法很多已不再适用. 因此，研究大数据环境下的机器学习算法成为学术界和产业界共同关注的话题.

本文主要分析和总结当前用于处理大数据的机器学习算法的研究现状. 并行是处理大数据的主流方法，本文还单独介绍一些并行算法，并引出大数据环境下机器学习研究所面临的问题.

2 大数据的相关知识介绍

2.1 大数据定义

有关大数据的定义有多种. 一个狭义的定义：大数据是指不能装载进计算机内存储器的数据. 尽管这是一个非正式的定义，但易理解，因为每台电脑都有一个大到不能装载进内存的数据集. 李国杰等[3]对大数据的定义为：一般意义上，大数据是指无法在可容忍的时间内用传统 IT 技术和软硬件工具对其进行感知、获取、管理、处理和服务的数据集合.

2.2 大数据特点

大数据有多方面的特点，从最开始的 3V 模型到目前扩展的 4V 模型就是以大数据的特点命名的. Laney 的 3V 模型包括体积(Volume)，速度(Velocity)和多样性(Variety)；4V 模型中的第 4 个 V 有多种解释，如变化性(Variability)，虚拟化(Virtual)或价值(Value). 针对这些特点，王飞跃[4] 认为在大数据时代知识解析、机器智能与人类智能协调工作及智能分析系统将会扮演重要角色，人们需要一种智能分析接口将人类与计算机世界连接，否则将被淹没在大数据的洪流中.

总之，大数据问题是目前学术界和产业界共同关注的挑战性问题. 伴随着大数据的采集、传输、处理和应用的相关技术就是大数据处理技术，是系列使用非传统的工具来处理大量的结构化、半结构化和非结构化数据，从而获得分析和预测结果的一系列数据处理技术[5].

3 大数据环境下的机器学习理论

随着大数据时代的到来，大数据逐渐成为学术界和产业界的热点，已在很多技术和行业广泛应用，从大规模数据库到商业智能和数据挖掘应用；从搜索引擎到推荐系统；推荐最新的语音识别、翻译等. 大数据算法的设计、分析和工程涉及很多方面，包括大规模并行计算、流算法、云技术等. 由于大数据存在复杂、高维、多变等特性，如何从真实、凌乱、无模式和复杂的大数据中挖掘出人类感兴趣的知识，迫切需要更深刻的机器学习理论进行指导.

传统机器学习的问题主要包括如下 4 个方面[5]：1)理解并且模拟人类的学习过程；2)针对计

算机系统和人类用户之间的自然语言接口的研究;3)针对不完全的信息进行推理的能力,即自动规划问题;4)构造可发现新事物的程序.

传统机器学习面临的一个新挑战是如何处理大数据. 目前,包含大规模数据的机器学习问题是普遍存在的,但是,由于现有的许多机器学习算法是基于内存的,大数据却无法装载进计算机内存,故现有的诸多算法不能处理大数据. 如何提出新的机器学习算法以适应大数据处理的需求,是大数据时代的研究热点方向之一.

4　大数据环境下的机器学习算法

4.1　大数据分治策略与抽样

分治策略是一种处理大数据问题的计算范例,尤其是近来在分布式和并行计算有很大发展的情况下,分治策略显得尤为重要.

一般来说,数据中不同样本对学习结果的重要程度也不相同. 一些冗余和噪音数据不仅造成大量的存储耗费,降低学习算法运行效率,而且还会影响学习精度. 因此更倾向于依据一定的性能标准(如保持样本的分布、拓扑结构及保持分类精度等)选择代表性样本形成原样本空间的一个子集,之后在这个子集上构造学习方法,完成学习任务. 这样能在不降低甚至提高某方面性能的基础上,最大限度地降低时间空间的耗费. 在大数据的背景下,样本选取的需求更迫切. 但已有的大部分样本选取方法只适用于较小规模的数据集,如传统的压缩最近邻(Condensed Nearest Neighbor, CNN)[6]、约减最近邻(Reduced Nearest Neighbor, RNN)[7]、编辑最近邻(Edited Nearest Neighbor, ENN)[8],它们的核心概念都是最小一致子集,而要找到这个子集需逐一测试样本,并且对子集的初始化及样本加入子集的顺序很敏感. 文献[9]基于局部几何和概率分布来选择分类边缘和边界样本,保持原有数据的空间信息,但需计算每个样本的 k 近邻. 文献[10]和文献[11]在CNN的基础上为近邻决策规则提出一种快速压缩最近邻算法(Fast CNN, FCNN),倾向于选择分类边界的样本.

Jordan[12]提出一些关于大数据的统计推理的方法. 当用分治算法来处理统计推理问题时,需从庞大的数据集中获取置信区间. Bootstrap 理论是通过重新采样数据来获取评估值的波动,进而获取置信区间,然而这对大数据是不可行的. 数据的不完全抽样会导致错误范围上的波动,必须进行更正以提供校准的统计推理. Jordan 提出一种程序——Bag of Little Bootstraps,可避开这一问题,不仅继承 Bootstrap 的理论性质,并且有许多计算上的优势. Jordan 讨论的另一个问题是大规模矩阵计算. 分治策略是一种启发式策略,在实际应用中有较好效果,但当试图描述分治算法的统计性能时,新的理论问题就会出现. 基于此,Jordan 给出基于随机矩阵的浓度定理的理论支持.

综上所述,数据分治与并行处理策略是大数据处理的基本策略,但目前的分治与并行处理策略较少利用大数据的分布知识,且影响大数据处理的负载均衡与计算效率. 如何学习大数据的分布知识用于优化负载均衡是一个亟待解决的问题.

4.2　大数据特征选择

在数据挖掘、文档分类和多媒体索引等新兴领域中,所面临的数据对象往往是大数据集,其中包含的属性数和记录数都很大,导致处理算法的执行效率低下. 通过属性选择可剔除无关属性,增加分析任务的有效性,从而提高模型精度,减少运行时间.

大数据处理的一个巨大挑战是如何处理高维、稀疏数据. 大数据环境下,网络流量、通讯记录、大规模社会网络产生大量、多方面的高维数据,张量(如多维数组)表示法提供一种大数据的自然表示. 故张量分解成为一种重要的汇总和分析工具. Kolda[13]提出一种内存使用高效的 Tucker 分解方法(Memory-Efficient Tucker Decomposition, MET),用于解决传统的张量分解算法无法解决的时间和空间利用问题. MET 在分解的过程中基于可用内存自适应选择正确的执行策略. 该算法在利用可用内存的前提下最大化计算速度. MET 避免处理在计算过程中产生的大量的零星中间结果,自适应选择操作顺序,不仅消除中间溢出问题,而且在不减少精确度的前提下节省内存. 另一方面,Wahba[14]探讨对此涉及到离散、含噪声、不完全相异数据统计/机器学习模型的两种方法——监督、半监督或无监督的机器学习方法. 这两种方法是正则化核估计(Regularized Kernel Estimation, RKE)和鲁棒流形展开(Robust Manifold Unfolding, RMU). 这些方法使用训练集中对象之间相异的信息,得到一个非负的低阶正定矩阵,用于将对象嵌入到一个低维欧几里德空间,其坐标可被用作各种学习模式中的属性. 同样,大多数在线学习研究需访问训练实例的所有属性/特征,对高维数据实例或当获得完整的属性/特征集合很昂贵时,这样的经典场景并不总是适合实际应用. 为突破此限制,Hoi 等[15]通过研究稀疏正则化和截断技

术,解决在线特征选择如何利用一个小的和固定数目的活跃特征来准确预测,提出一个有效的算法,评估所提出算法在一些公共数据集上在线特征选择的经验性能,并证明用在线特征选择技术解决现实世界大数据挖掘问题比一些著名的批处理特征选择算法具有更显著的可扩展性.

传统的自组织映射(Self-organizing Map, SOM)可用于特征提取,但当数据集较大时,SOM 就存在速度慢的缺点,Sagheer 等[16]提出一种快速自组织映射(Fast SOM, FSOM)主要用于解决该问题. 该方法的主要思想是:大数据的主要信息主要分布于特征空间的某一区域,如果能找到这些区域并直接在这些区域提取信息,而不是在整个数据空间提取信息,则能大幅度减少时间.

Anaraki 等[17]提出一种带阈值的基于模糊下近似的模糊粗糙集特征选择方法(Fuzzy Lower-Approximation-Based Fuzzy Rough Set Feature Selection with Threshold, T-L-FRFS). 该方法加入一个阈值来限制 QuickReduct 选取特征的数量,与传统的基于模糊下近似的模糊粗糙集特征选择方法(L-FRFS)相比,该方法减少选取特征的数量. 实验结果也证明该方法可提高特征提取准确率,减少运行时间.

Quevedo 等[18]基于输入变量的有用性,采用经典技术的简单组合,如相关性和正交性,提出一种输入变量排名算法,用于大数据降维和特征提取,取得良好效果.

Gheyas 等[19]结合模拟退火算法、遗传算法、贪心算法及神经网络算法的优点,提出一种模拟退火和遗传算法(Simulated Annealing and Genetic Algorithm, SAGA)混合算法用于解决选择最优化特征子集的 NP 时间问题. 实验结果表明,该算法能得到更好的最优化特征子集,并能降低时间复杂度. Gheyas 等在论文结尾之处指出,很少有一种单一算法能解决所有问题,但如果能将一些算法结合,则能较好地弥补算法相互间的不足.

Pal 等[20]提出一种基于 SVM 的用于分类的特征选择方法,SVM 分类算法的准确度与数据集的特征数和数据集大小有关,因此,在分类前对数据进行特征选择有利于提高分类准确度. 然而特征选择方法也很重要,不同特征选择方法所选出来的特征区别很大.

Sun 等[21]提出一种用于分类的特征选择算法. 该算法利用局部学习理论首先将复杂的非线性问题转换为一组线性问题,然后在最大间隔的框架下学习特征关联性. 该算法基于机器学习和数值分析技术等,不需假设有关数据分布. 由于不需任何启发搜索,该算法在含有大量不相关特征的数据集上也具有较好性能.

Hua 等[22]对比一些现有的特征选择方法,提出一种特征标签分布式模型,在这个模型中可用特征的数量与它们在真实数据中是一样的,通过它们在一些高维数据上的测试结果显示,不同的特征选择算法在不同的模型条件下性能不一样,它们与样本数量、全局和异构标记的数量有关. 该模型不仅局限用于论文中所提到的特征选择算法,而且也适用于其他的分类算法和特征选择算法.

Song 等[23]研究降维技术在轨迹聚类中的作用,运用 3 种主流降维方法 SVD、RP 和 PCA,结合最通用的轨迹聚类算法,可在计算时间和本地流程模型的适合程度上提高算法性能.

综上所述,由于大数据存在复杂、高维、多变等特性,如何采用降维和特征选择技术以降低大数据处理难度,是大数据特征选择技术迫切需要解决的问题.

4.3 大数据分类

有监督学习(分类)面临的一个新挑战是如何处理大数据. 目前包含大规模数据的分类问题是普遍存在的,但是传统分类算法不能处理大数据.

1)支持向量机分类. 传统统计机器学习方法用于大数据分类有两大瓶颈问题:(1)计算密集型,几乎不能用于大规模数据集;(2)鲁棒和非参数的置信区间的拟合模型的预测往往是未知的. 针对上述问题,Lau 等[24]为 SVM 提出一种在线学习算法,用于处理按顺序逐渐提供输入数据的分类问题. 该算法速度更快,所用支持向量个数更少,并具有更优的泛化能力. Laskov 等[25]提出一种快速、数值稳定和鲁棒的增量支持向量机学习方法.

Huang 等[26]提出一种大边缘分类器 M^4,与其他大边缘分类器或局部地或全局地构建分离超平面不同,该模型能局部和全局地学习判定边界. SVM 和极大极小概率机(Minimax Probability Machine, MPM)与该模型具有密切联系,该模型具有重要理论意义. 进一步地,M^4 的最优化问题可在多项式时间内解决.

针对大规模数据的分类问题,在增量核主成分分析(Incremental Kernel PCA)和基于共轭梯度的最小二乘支持向量机(Least Squares SVM, LS-SVM)算法基础之上,Kim 等[27]提出适用于大数据的特征提取和分类算法. 该算法所需内存较少,无需存储较大

矩阵,可更好地解决大规模数据分类问题.

2)决策树分类. 传统决策树(Decision Tree)作为一种经典的分类学习算法,对大数据处理存在内存开销过大的问题. Franco-Arcega 等[28]提出一种从大规模数据中构造决策树的方法,解决当前算法中的一些限制条件,可利用所有的训练集数据,但不需将它们都保存在内存中. 实验表明,该方法比目前的决策树算法在大规模问题上计算速度更快. Yang 等[29]提出一种增量优化的快速决策树算法(Incrementally Optimized Very Fast Decision Tree, iOVFDT)用于处理带有噪音的大数据,与传统的挖掘大数据的决策树算法相比,该算法的主要优势是实时挖掘能力,这使得当移动数据流是无限时,它能存储完整的数据用于再训练决策模型. 此模型的优点在于在数据流包含有噪音时,能阻止生成决策树大小的爆炸性增长及预测精度的恶化. 该模型即使是在含有噪音的数据中,也能生成紧凑的决策树,并有较高的预测精度. Ben-Haim 等[30]提出一种构建决策树分类器的算法. 该算法在分布式环境中运行,适用于大数据集和流数据,与串行决策树相比,该方法在精度误差近似的前提下能提高效率.

3)神经网络与极端学习机(Extreme Learning Machine, ELM). 传统前馈神经网络一般采用梯度下降算法调整权值参数,学习速度慢、泛化性能差等问题是制约前馈神经网络应用的瓶颈. 最近 Huang 等[31]摒弃梯度下降算法的迭代调整策略,提出 ELM. 该方法随机赋值单隐层神经网络的输入权值和偏差项,并通过一步计算即可解析求出网络的输出权值. 相比于传统前馈神经网络训练算法需经多次迭代调整才可最终确定网络权值,ELM 的训练速度获得较显著提升.

然而,由于计算能力和复杂性的限制,在大数据上训练出单一 ELM 是一个困难的问题. 解决这一困难通常有两种途径:1)训练基于分治策略的 ELM[32];2)在训练单一 ELM 中引入并行机制[33]. 文献表明单一 ELM 有着很强的函数逼近能力[34-35],可否将这种逼近能力延拓到基于分治策略的 ELM 是衡量 ELM 是否适用于大数据学习的一个关键指标. 尚未见到基于分治策略的 ELM 逼近的研究报道,一些相关的研究还包括有效性学习[36].

4)应用领域的分类算法. 除此之外,在一些应用领域,也有针对大数据的分类算法提出. 在计算机辅助诊断领域,机器学习广泛应用于帮助医学专家从已诊断案例中获取先验知识,但大量的已诊断样本很难获取. Li 等[37]提出一种半监督的学习算法——基于随机森林的协同训练(Co-training Based on Random Forest, Co-forest),用来估计未诊断样本的标记自信度,能较易得出先验知识. 该方法在基准数据集上得到较好结果. 针对大规模图像数据集的分类性能问题,Lin 等[38]提出在特征提取和分类器训练方面提高效率. 对于特征提取,利用 Hadoop 架构,在几百个 Mapper 上并行计算. 对于训练 SVM,提出并行平均随机梯度下降算法(Parallel Averaging Stochastic Gradient Descent, ASGD),可处理具有 120 万个图像、1 000 类的图像数据,并具有较快的收敛速度.

另外,中文网页标记数据稀缺,英文网页标记数据较丰富,Ling 等[39]用英文网页标记信息来解决跨语言分类问题,提出基于信息瓶颈(Information Bottleneck)的方法. 该方法首先将中文翻译成英文,然后将所有网页通过一个只允许有限信息通过的信息瓶颈来编码. 该方法可使跨语言分类更准确,较显著提高一些已有的监督与半监督分类器的准确率.

综上所述,传统机器学习的分类方法很难直接运用到大数据环境下,不同的分类算法都面临着大数据环境的挑战,针对不同分类算法如何研究并行或改进策略成为大数据环境下分类学习算法研究的主要方向.

4.4　大数据聚类

聚类学习是最早被用于模式识别及数据挖掘任务的方法之一,并且被用来研究各种应用中的大数据库,因此用于大数据的聚类算法受到越来越多的关注. Havens 等[40]对比 3 种扩展的模糊 c 均值(FCM)聚类算法对于大数据的执行效率. 具体而言,这 3 种方法分别基于:1)取样后进行非迭代扩展;2)连续通过数据子集的增量技术;3)提供基于抽样的估计的核模糊 c 均值算法. Havens 等用可装载的数据集和 VL 数据集来进行数值型实验,这些实验进行如下对比:时间复杂度、空间复杂度、速度、处理装载数据的批量 FCM 的近似质量、对划分和地面实况间匹配的评估. 实验结果显示,随机取样可扩展 FCM(Random Sampling Plus Extension FCM),位减少 FCM(Bit-Reduced FCM)及近似核 FCM(Approximate Kernel FCM)都是较好的选择,都近似于 FCM. 最后,Havens 等展示针对含有 50 亿对象的数据集的大数据算法,并就如何使用不同的大数据 FCM 聚类策略提出一系列建议.

另一方面,随着数据体积的增大,I/O瓶颈就变成数据分析的一个重要问题. 数据压缩能起到缓解作用. 以 K-means 为例,Xue 等[41]提出一种压缩感

知性能提升模型用于大数据聚类. 该模型定量分析整个计算过程中与压缩有关的诸多因素的影响. 在有上百个计算核的集群上对大到 1.114 TB 的 10 维数据进行聚类实验,实验结果证明使用压缩能改善 I/O 性能,并且该模型能有效决定何时如何使用压缩来改善大数据分析中的 I/O 性能. 针对分布式聚类、流数据聚类,Hall 等[42] 研究二次抽样方法以提高聚类算法的可扩展性. 实验表明,人们可构造一个好的模型而不必知道所有的数据,如果需要,修改后的算法可应用于 TB 级或更多的数据.

为实现数据的大规模并行处理,MapReduce 模型成为学术界和工业界最为流行的工具. 从数据库角度看,MapReduce 是一种简单但强大的执行引擎,可与其他数据存储和管理组件有效融合. 为解决大规模数据分析难题,Zhao 等[43] 提出基于 MapReduce 的 *K*-means 算法,在 speedup、sizeup、scaleup 这 3 个指标上获得较好的并行性能. Papadimitriou 等[44] 给出一种利用 MapReduce 模型实现协同聚类(Co-clustering)的系统框架——分布式协同聚类框架(Distributed Co-clustering, DisCo),并引入分布式数据预处理、协同聚类等方法,在 Hadoop 上实现该系统. 实验结果证明 DisCo 具有良好的可扩展性、高执行效率,能处理几百 GB 数据. Zhang 等[45] 针对大规模数据分析需求,提出一种基于 MapReduce 的并行 KNN 算法,通过在一系列真实和模拟数据集上的实验证明该算法在大规模数据处理方面具有良好的执行效率和可扩展性. Ferreira 等[46] 给出一种利用 MapReduce 开展大规模数据聚类的方法. 主要针对问题有:1)如何将 I/O 开销最小化;2)如何在处理节点之间降低通信开销. 上述两个方面成为系统性能瓶颈. 所提出的 BoW(Best of both Worlds)方法能自动发现系统瓶颈并选择应对策略. 总体来说,其贡献在于:1)提出 BoW,并设计开销函数,动态选择最优策略;2)实验证明 BoW 有明显优势. Havens 等[47] 分析在大数据上运行 *C*-mean 的困难,指出模糊技术在处理大数据上的有效性,研究抽样和增量在大数据上运行 *C*-mean 的作用.

随着信息技术的迅猛发展,聚类所面临的不仅是数据量越来越大的问题,更重要的还是数据的高维问题. 由于数据来源的丰富多样,图文声像甚至视频都逐渐成为聚类处理的目标对象,这些特殊对象的属性信息往往要从数百个甚至成千上万个方面来表现,其每个属性都成为数据对象的一个维,对高维数据的聚类分析,已成为众多领域研究方向之一. 高维数据的聚类方法包括基于降维的聚类[48]、子空间聚类[49-50]、基于图的聚类[51] 等. 在很多需处理高维数据的应用领域,降维是常用的方法之一. 直观地讲,降维就是通过把数据点映射到更低维的空间上以寻求数据的紧凑表示的一种技术,这种低维空间的紧凑表示将有利于对数据的进一步处理. 降维从一般意义上代表着数据信息的损失,如何保证降维而不损失聚类性能是传统数据领域仍在研究和讨论的问题.

综上所述,经典的聚类算法在大数据环境下面临诸如数据量大、数据体积过大、复杂度高等众多挑战,如何并行或改进现有聚类算法,进而提出新的聚类算法成为研究关键.

4.5 大数据关联分析

关联分析的研究源自于 Apriori 算法[52],Apriori 性质要求频繁模式的子模式必须也是频繁的. 基于这一启发,出现一系列的类 Apriori 算法:AprioriAll, AprioriSome, DynamicSome[53],广义序列模式(Generalized Sequential Pattern, GSP)[54] 及基于等价类的序列模式发现(Sequential Pattern Discovery Using Equivalence Classes, SPADE)[55]. 后来,研究者又提出一系列基于数据投影的算法,包括频繁模式投影序列模式挖掘算法(Frequent Pattern-Projected Sequential Pattern Mining, FreeSpan)[56] 和前缀投影序列模式挖掘算法(Prefix-Projected Sequential Pattern Mining, PrefixSpan)[57]. SPADE 是基于格点的算法. 基于内存索引的序列模式挖掘算法(Memory Indexing for Sequential Pattern Mining, MEMISP)[58] 是基于内存索引的方法. 而基于正则表达式约束的序列模式挖掘算法(Sequential Pattern Mining with Regular Expression Constraints, SPIRIT)[59] 通过使用正则表达式将约束整合到一起.

解决大数据的关联分析主要有两种途径:并行和增量.

在并行方面,Li 等[60] 提出一种基于 MapReduce 的并行 Apriori 算法. Apriori 算法最主要的操作为产生候选项集,该算法将产生候选项集的过程并行化,提高运行效率,并具有良好的加速比和伸缩性.

增量方面主要体现在序列模式挖掘上. 在文献[61] 中,作者介绍基于广义序列模式(Generalized Sequential Pattern, GSP)和基于 GSP 的频繁序列挖掘算法(Mining Frequent Sequences, MFS)的增量挖掘算法 GSP + 和 MFS +. 文献[62] 中提出基于 SPADE 的增量序列挖掘算法(Incremental Sequence Mining, ISM),该方法不仅可在数据库更新时维持

频繁序列,还提供一个用户交互接口以方便用户修改约束,如最小支持度等. ISM 只考虑序列追加,而文献[63]所提的增量频繁序列挖掘算法(Incremental Frequent Sequences Mining, ISE)还考虑插入新序列的情况. 然而,ISE 只考虑对频繁序列的后缀的扩充. 文献[64]中提出的增量更新序列算法(Incrementally Updating Sequences, IUS)对旧的频繁序列的前缀和后缀都可进行扩充. 文献[65]中算法可在维持序列模式的同时从数据库中删除一些记录. 文献[66]中提出一个实验性方法——性能与差异均衡算法(Tradeoff between Performance and Difference, TPD),来确定何时更新序列模式. 以上增量算法都为提高算法效果,在大数据上的运行效果有待进一步验证.

4.6　大数据并行算法

如何把传统机器学习算法运用到大数据环境中,一个典型策略是对现有学习算法并行化. 例如,图形处理器(Graphic Processing Unit,GPU)平台从并行上得到较显著的性能提升. 这些 GPU 平台由于采用并行架构,使用并行编程方法,使得计算能力呈几何级数增长[67]. Luo 等[68]提出一种非平凡的策略用来并行处理一系列数据挖掘与机器学习问题,包括一类分类 SVM 和两类分类 SVM、非负最小二乘问题、L1 正则化回归问题. 由此得到的乘法算法,可直接在如 MapReduce 和通用并行计算架构(Compute Unified Device Architecture, CUDA)的并行计算环境中实现. 在大数据分类和聚类学习当中,MapReduce 框架被用于并行化传统的机器学习算法以适应大数据处理的需求. Shim[69]在 MapReduce 框架下,讨论如何设计高效的 MapReduce 算法,对当前一些基于 MapReduce 的机器学习和数据挖掘算法归纳总结,以便进行大数据的分析. Zhang 等[70]提出一种大数据挖掘技术,即利用 MapReduce 实现并行的基于粗糙集的知识获取算法,还提出下一步的研究方向,即集中于基于并行技术的粗糙集算法处理非结构化数据.

Hefeeda 等[71]提出一种近似算法使基于核的机器学习算法可有效处理大规模数据集. 当前的基于核的机器学习算法由于需计算核矩阵而面临着可伸缩性问题,这是因为计算核矩阵需 $O(n^2)$ 的时间和空间复杂度. 该算法计算核矩阵时不仅大幅降低计算和内存开销,而且没有明显影响结果的精确度.

此外,由于传统的提升算法本身具有串行特点,不易获得好的扩展性,因此,需提出并行的提升算法来高效处理大规模数据. Palit 等[72]提出 2 种并行提升算法,即 ADABOOST. PL 和 LOGITBOOST. PL,它们可使多个计算节点同时参与计算并且可构造出一个提升集成分类器. 该方法通过利用 MapReduce 框架来实现,在合成数据和真实数据集上的实验表明其在分类准确率、加速比和放大率方面都取得较好结果.

Kaiser 等[73]还利用 MapReduce 框架分布式实现一系列核函数学习机训练,该方法适用于基于核的分类和回归. Kaiser 还介绍一种扩展版的区域到点建模方法,适应来自空间区域的大量数据.

Yan[74]考虑潜在狄利克雷分配模型(Latent Dirichlet Allocation, LDA)的两种推理方法——塌缩吉布斯采样(Collapsed Gibbs Sampling, CGS)和塌缩变分贝叶斯推理(Collapsed Variational Bayesian, CVB)在 GPU 上的并行化问题. 为解决 GPU 上的有限内存限制问题,Yan 提出一种能有效降低内存开销的数据划分方案. 这种划分方案也能平衡多重处理器的计算开销,并能避免内存访问冲突,使用数据流来处理超大的数据集.

针对异构云中进行大数据分析服务的并行化问题,Jung 等[75]提出最大覆盖装箱算法来决定系统中多少节点、哪些节点应该应用于大数据分析的并行执行. 该方法可使大数据进行分配,使得各个计算节点可同步结束计算,并且使数据块的传输可以和上一个块的计算重叠,从而节省时间. 实验表明,该方法比其他方法可提高大约 60% 的性能.

在分布式系统方面,中国科学院计算技术研究所智能信息处理重点实验室数据挖掘与机器学习组在 2008 年底与中国移动合作,开发完成分布式并行数据挖掘系统 PDMiner(Parallel Distributed Miner),这是中国最早的基于云计算平台的并行数据挖掘系统之一. 系统提供多种并行数据转换规则和并行数据挖掘算法,已用于中国移动通信企业 TB 级实际数据的挖掘,达到商用软件的精度[43,60,76-81].

Cheng 等[82]提出一个面向大规模可伸缩数据分析的可伸缩的分布式系统——广义线性聚合分布式引擎(Generalized Linear Aggregates Distributed Engine, GLADE). GLADE 通过用户自定义聚合(User-Defined Aggregate, UDA)接口并且在输入数据上有效运行来进行数据分析. 作者从两个方面来论证系统的有效性. 1)展示如何使用一系列分析功能来完成数据处理. 2)将 GLADE 与两种不同类型的系统对比:用 UDA 进行改良的关系型数据库(PostgreSQL)和 Map-Reduce(Hadoop). 然后从运行结果、伸缩性及运行时间上与不同类型的系统对比.

综上所述,并行策略是传统机器学习算法运用

于大数据的典型策略之一，并且在一定范围内取得一些进展，能处理一定量级的大数据. 如何研究高效的并行策略以高效处理大数据也是当今的研究热点之一.

5 结 束 语

大数据具有属性稀疏、超高维、高噪声、数据漂移、关系复杂等特点，导致传统机器学习算法难以有效处理和分析，为此，需在如下方面展开相应研究. 1)研究机器学习理论和方法，包括数据抽样和属性选择等大数据处理的基本技术，设计适合大数据特点的数据挖掘算法，以实现超高维、高稀疏的大数据中的知识发现. 2)研究适合大数据分布式处理的数据挖掘算法编程模型和分布式并行化执行机制，支持数据挖掘算法迭代、递归、集成、归并等复杂算法编程. 3)在 Hadoop、CUDA 等并行计算平台上，设计和实现复杂度低、并行性高的分布式并行化机器学习与数据挖掘算法.

致 谢 衷心感谢课题组的敖翔，马云龙，尚田丰，董智，韩硕，王浩成，余文超，金鑫，杜长营，吴新宇，程小虎，他们共同参与文献调研工作.

参 考 文 献

[1] Labrinidis A, Jagadish H V. Challenges and Opportunities with Big Data. Proc of the VLDB Endowment, 2012, 5(12): 2032 - 2033

[2] Bizer C, Boncz P, Brodie M L, *et al*. The Meaningful Use of Big Data: Four Perspectives - Four Challenges. ACM SIGMOD Record, 2012, 40(4): 56 - 60

[3] Li G J, Cheng X Q. Research Status and Scientific Thinking of Big Data. Bulletin of Chinese Academy of Sciences, 2012, 27(6): 647 - 657 (in Chinese)
(李国杰，程学旗. 大数据研究：未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考. 中国科学院院刊, 2012, 27(6): 647 - 657)

[4] Wang F Y. A Big-Data Perspective on AI: Newton, Merton, and Analytics Intelligence. IEEE Intelligent Systems, 2012, 27(5): 2 - 4

[5] Simon H A. Why Should Machines Learn? // Michalski R S, Carbonell J G, Mitchell T M, *et al*., eds. Machine Learning: An Artificial Intelligence Approach. Berlin, Germany: Springer, 1983: 25 - 37

[6] Hart P. The Condensed Nearest Neighbor Rule. IEEE Trans on Information Theory, 1968, 14(3): 515 - 516

[7] Gates G. The Reduced Nearest Neighbor Rule. IEEE Trans on Information Theory, 1972, 18(3): 431 - 433

[8] Brighton H, Mellish C. Advances in Instance Selection for Instance-Based Learning Algorithms. Data Mining and Knowledge Discovery, 2002, 6(2): 153 - 172

[9] Li Y H, Maguire L. Selecting Critical Patterns Based on Local Geometrical and Statistical Information. IEEE Trans on Pattern Analysis and Machine Intelligence, 2011, 33(6): 1189 - 1201

[10] Angiulli F. Fast Nearest Neighbor Condensation for Large Data Sets Classification. IEEE Trans on Knowledge and Data Engineering, 2007, 19(11): 1450 - 1464

[11] Angiulli F, Folino G. Distributed Nearest Neighbor-Based Condensation of Very Large Data Sets. IEEE Trans on Knowledge and Data Engineering, 2007, 19(12): 1593 - 1606

[12] Jordan M I. Divide-and-Conquer and Statistical Inference for Big Data // Proc of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012. DOI:10.1145/2339530.2339534

[13] Kolda T G, Sun J M. Scalable Tensor Decompositions for Multi-aspect Data Mining // Proc of the 8th IEEE International Conference on Data Mining. Pisa, Italy, 2008: 363 - 372

[14] Wahba G. Dissimilarity Data in Statistical Model Building and Machine Learning // Proc of the 5th International Congress of Chinese Mathematicians. Beijing, China, 2012: 785 - 809

[15] Hoi C H, Wang J L, Zhao P L, *et al*. Online Feature Selection for Mining Big Data // Proc of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications. Beijing, China, 2012: 93 - 100

[16] Sagheer A, Tsuruta N, Taniguchi R I, *et al*. Fast Feature Extraction Approach for Multi-dimension Feature Space Problems // Proc of the 18th International Conference on Pattern Recognition. Hong Kong, China, 2006, III: 417 - 420

[17] Anaraki J R, Eftekhari M. Improving Fuzzy-Rough Quick Reduct for Feature Selection // Proc of the 19th Iranian Conference on Electrical Engineering. Tehran, Iran, 2011: 1 - 6

[18] Quevedo J R, Bahamonde A, Luaces O. A Simple and Efficient Method for Variable Ranking according to Their Usefulness for Learning. Computational Statistics & Data Analysis, 2007, 52(1): 578 - 595

[19] Gheyas I A, Smith L S. Feature Subset Selection in Large Dimensionality Domains. Pattern Recognition, 2010, 43(1): 5 - 13

[20] Pal M, Foody G M. Feature Selection for Classification of Hyperspectral Data by SVM. IEEE Trans on Geoscience and Remote Sensing, 2010, 48(5): 2297 - 2307

[21] Sun Y J, Todorovic S, Goodison S. Local-Learning-Based Feature Selection for High-Dimensional Data Analysis. IEEE Trans on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1610 - 1626

[22] Hua J P, Tembe W D, Dougherty E R. Performance of Feature-Selection Methods in the Classification of High-Dimension Data. Pattern Recognition, 2009, 42(3): 409 - 424

[23] Song M, Yang H, Siadat S H, *et al*. A Comparative Study of Dimensionality Reduction Techniques to Enhance Trace Clustering Performances. Expert Systems with Applications, 2013, 40(9):

3722 – 3737

[24] Lau K W, Wu Q H. Online Training of Support Vector Classifier. Pattern Recognition, 2003, 36(8): 1913 – 1920

[25] Laskov P, Gehl C, Krüger S, *et al*. Incremental Support Vector Learning: Analysis, Implementation and Applications. Journal of Machine Learning Research, 2006, 7: 1909 – 1936

[26] Huang K, Yang H, King L, *et al*. Maxi-Min Margin Machine: Learning Large Margin Classifiers Locally and Globally. IEEE Trans on Neural Networks, 2008, 19(2): 260 – 272

[27] Kim B J. A Classifier for Big Data // Proc of the 6th International Conference on Convergence and Hybrid Information Technology. Daejeon, Republic of Korea, 2012: 505 – 512

[28] Franco-Arcega A, Carrasco-Ochoa J A, Sánchez-Díaz G, *et al*. Building Fast Decision Trees from Large Training Sets. Intelligent Data Analysis, 2012, 16(4): 649 – 664

[29] Hang Y, Fong S. Incrementally Optimized Decision Tree for Noisy Big Data // Proc of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications. Beijing, China, 2012: 36 – 44

[30] Ben-Haim Y, Tom-Tov E. A Streaming Parallel Decision Tree Algorithm. Journal of Machine Learning Research, 2010, 11: 849 – 872

[31] Huang G B, Zhu Q Y, Siew C K. Extreme Learning Machine: Theory and Applications. Neurocomputing, 2006, 70(1/2/3): 489 – 501

[32] Liu N, Wang H. Ensemble Based Extreme Learning Machine. IEEE Signal Processing Letters, 2010, 17(8): 754 – 757

[33] He Q, Shang T F, Zhuang F Z, *et al*. Parallel Extreme Learning Machine for Regression Based on MapReduce. Neurocomputing, 2013, 102: 52 – 58

[34] Zhang R, Lan Y, Huang G B, *et al*. Universal Approximation of Extreme Learning Machine with Adaptive Growth of Hidden Nodes. IEEE Trans on Neural Networks and Learning Systems, 2012, 23(2): 365 – 371

[35] Rong H J, Huang G B, Sundararajan N, *et al*. Online Sequential Fuzzy Extreme Learning Machine for Function Approximation and Classification Problems. IEEE Trans on Systems, Man and Cybernetics, 2009, 39(4): 1067 – 1072

[36] Yang Y M, Wang X N, Yuan X F. Bidirectional Extreme Learning Machine for Regression Problem and Its Learning Effectiveness. IEEE Trans on Neural Networks and Learning Systems, 2012, 23(9): 1498 – 1505

[37] Li M, Zhou Z H. Improve Computer-Aided Diagnosis with Machine Learning Techniques Using Undiagnosed Samples. IEEE Trans on Systems, Man and Cybernetics, 2007, 37(6): 1088 – 1098

[38] Lin Y Q, Lü F J, Zhu S H, *et al*. Large-Scale Image Classification: Fast Feature Extraction and SVM Training // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA, 2011: 1689 – 1696

[39] Ling X, Xue G R, Dai W Y, *et al*. Can Chinese Web Pages Be Classified with English Data Source? // Proc of the 17th International Conference on World Wide Web. Beijing, China, 2008: 969 – 978

[40] Havens T C, Bezdek J C, Leckie C, *et al*. Fuzzy *c*-means Algorithms for Very Large Data. IEEE Trans on Fuzzy Systems, 2012, 20(6): 1130 – 1146

[41] Xue Z H, Shen G, Li J H, *et al*. Compression-Aware I/O Performance Analysis for Big Data Clustering // Proc of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications. Beijing, China, 2012: 45 – 52

[42] Hall L O. Exploring Big Data with Scalable Soft Clustering // Proc of the 6th International Conference on Soft Methods in Probability and Statistics. Konstanz, Germany, 2012: 11 – 15

[43] Zhao W Z, Ma H F, He Q. Parallel *k*-means Clustering Based on MapReduce // Proc of the 1st International Conference on Cloud Computing and Big Data. Beijing, China, 2009: 674 – 679

[44] Papadimitriou S, Sun J M. DisCo: Distributed Co-clustering with MapReduce: A Case Study towards Petabyte-Scale End-to-End Mining // Proc of the 8th IEEE International Conference on Data Mining. Pisa, Italy, 2008: 512 – 521

[45] Zhang C, Li F F, Jeffrey J. Efficient Parallel *k*NN Joins for Large Data in MapReduce // Proc of the 15th International Conference on Extending Database Technology. Berlin, Germany, 2012: 38 – 49

[46] Ferreira C R L, Junior T C, Traina A J M, *et al*. Clustering Very Large Multi-dimensional Datasets with MapReduce // Proc of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA, 2011: 690 – 698

[47] Havens T C, Chitla R, Jain A K, *et al*. Speedup of Fuzzy and Possibilistic Kernel *c*-means for Large-Scale Clustering // Proc of the IEEE International Conference on Fuzzy Systems. Taipei, China, 2011: 463 – 470

[48] Niu D L, Dy J G, Jordan M I. Dimensionality Reduction for Spectral Clustering // Proc of the 14th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, USA, 2011: 552 – 560

[49] Kriegel H P, Kröger P, Zimek A. Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering. ACM Trans on Knowledge Discovery from Data, 2009, 3(1): 1 – 58

[50] Vidal R. Subspace Clustering. IEEE Trans on Signal Processing, 2011, 28(2): 52 – 68

[51] Zhou Y, Cheng H, Yu J X. Graph Clustering Based on Structural/Attribute Similarities. Proc of the VLDB Endowment, 2009, 2(1): 718 – 729

[52] Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules in Large Databases // Proc of the 20th International Conference on Very Large Data Bases. Santiago de Chile, Chile, 1994: 487 – 499

[53] Agrawal R, Srikant R. Mining Sequential Patterns // Proc of the 11th International Conference on Data Engineering. Taipei, China, 1995: 3 – 14

[54] Srikanth R, Agrawal R. Mining Sequential Patterns: Generalizations and Performance Improvements // Proc of the 5th International Conference on Extending Database Technology: Advances in Database Technology. Avignon, France, 1996: 3 – 17

[55] Zaki M J. SPADE: An Efficient Algorithm for Mining Frequent Se-

quences. Machine Learning, 2001, 42(1/2): 31 - 60

[56] Han J W, Kamber M, Pei J. Data Mining Concepts and Techniques. 2nd Edition. New York, USA: Morgan Kaufmann, 2006

[57] Pei J, Han J W, Pinto H, *et al*. Prefixspan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth // Proc of the 17th International Conference on Data Engineering. Heidelberg, Germany, 2001: 215 - 224

[58] Lin M Y, Lee S Y. Fast Discovery of Sequential Patterns by Memory Indexing // Proc of the 4th International Conference on Data Warehousing and Knowledge Discovery. Aix-en-Provence, France, 2002: 150 - 160

[59] Garofalakis M N, Rastogi R, Shim K. Spirit: Sequential Pattern Mining with Regular Expression Constraints // Proc of the 25th International Conference on Very Large Data Bases. Edinburgh, Scotland, 1999: 223 - 234

[60] Li N, Zeng L, He Q, *et al*. Parallel Implementation of Apriori Algorithm Based on MapReduce // Proc of the 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing. Kyoto, Japan, 2012: 191 - 200

[61] Zhang M H, Kao B, Cheung D W, *et al*. Efficient Algorithms for Incremental Update of Frequent Sequences // Proc of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Taipei, China, 2002: 186 - 197

[62] Parthasarathy S, Zaki M J, Ogihara M, *et al*. Incremental and Interactive Sequence Mining // Proc of the 8th International Conference on Information and Knowledge Management. Kansas City, USA, 1999: 251 - 258

[63] Masseglia F, Poncelet P, Teisseire M. Incremental Mining of Sequential Patterns in Large Databases. Data & Knowledge Engineering, 2003, 46(1): 97 - 121

[64] Zheng Q G, Xu K, Ma S L, *et al*. The Algorithms of Updating Sequential Patterns [EB/OL]. [2013 - 05 - 20]. http://arxiv.org/ftp/cs/papers/0203/0203027.pdf

[65] Wang C Y, Hong T P, Tseng S S. Maintenance of Sequential Patterns for Record Deletion // Proc of the IEEE International Conference on Data Mining. San Jose, USA, 2001: 536 - 541

[66] Zheng Q G, Xu K, Ma S L. When to Update the Sequential Patterns of Stream Data? // Proc of the 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Seoul, Republic of Korea, 2003: 545 - 550

[67] Upadhyaya S R. Parallel Approaches to Machine Learning - A Comprehensive Survey. Journal of Parallel and Distributed Computing, 2013, 73(3): 284 - 292

[68] Luo D J, Ding C, Huang H. Parallelization with Multiplicative Algorithms for Big Data Mining // Proc of the 12th IEEE International Conference on Data Mining. Brussels, Belgium, 2012: 489 - 498

[69] Shim K. MapReduce Algorithms for Big Data Analysis. Proc of the VLDB Endowment, 2012, 5(12): 2016 - 2017

[70] Zhang J B, Li T R, Pan Y. Parallel Rough Set Based Knowledge Acquisition Using MapReduce from Big Data // Proc of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications. Beijing, China, 2012: 20 - 27

[71] Hefeeda M, Gao F, Abd-Almageed W. Distributed Approximate Spectral Clustering for Large-Scale Datasets // Proc of the 21st International ACM Symposium on High-Performance Parallel and Distributed Computing. Delft, the Netherlands, 2012: 223 - 234

[72] Palit I, Reddy C K. Scalable and Parallel Boosting with MapReduce. IEEE Trans on Knowledge and Data Engineering, 2012, 24(10): 1904 - 1916

[73] Kaiser C, Pozdnoukhov A. Enabling Real-Time City Sensing with Kernel Stream Oracles and MapReduce. Pervasive and Mobile Computing, 2013, 9(5): 708 - 721

[74] Yan F, Xu N Y, Qi Y. Parallel Inference for Latent Dirichlet Allocation on Graphics Processing Units // Proc of the 22nd Annual Conference on Neural Information Processing Systems. Whistler, Canada, 2009: 2134 - 2142

[75] Jung G, Gnanasambandam N, Mukherjee T. Synchronous Parallel Processing of Big-Data Analytics Services to Optimize Performance in Federated Clouds // Proc of the 5th IEEE International Conference on Cloud Computing. Hawaii, USA, 2012: 811 - 818

[76] He Q, Zhuang F Z, Li J C, *et al*. Parallel Implementation of Classification Algorithms Based on MapReduce // Proc of the 5th International Conference on Rough Set and Knowledge Technology. Beijing, China, 2010: 655 - 662

[77] He Q, Tan Q, Ma X D, *et al*. The High-Activity Parallel Implementation of Data Preprocessing Based on MapReduce // Proc of the 5th International Conference on Rough Set and Knowledge Technology. Beijing, China, 2010: 646 - 654

[78] He Q, Wang Q, Du C Y, *et al*. A Parallel Hyper-Surface Classifier for High Dimensional Data // Proc of the 3rd International Symposium on Knowledge Acquisition and Modeling. Wuhan, China, 2010: 338 - 343

[79] He Q, Ma Y L, Wang Q, *et al*. Parallel Outlier Detection Using KD-Tree Based on MapReduce // Proc of the 3rd International Conference on Cloud Computing Technology and Science. Athens, Greece, 2011: 75 - 80

[80] He Q, Wang Q, Zhuang F Z, *et al*. Parallel CLARANS Clustering Based on MapReduce. Energy Procedia, 2011, 13: 3269 - 3279

[81] Tan Q, He Q, Shi Z Z. Parallel Max-Min Ant System Using MapReduce // Proc of the 3rd International Conference on Swarm Intelligence. Shenzhen, China, 2012: 182 - 189

[82] Cheng Y, Qin C J, Rusu F. GLADE: Big Data Analytics Made Easy // Proc of the ACM SIGMOD International Conference on Management of Data. Scottsdale, USA, 2012: 697 - 700