DOI:10.13196/j.cims.2019.03.001

基于人工智能技术的大数据分析方法研究进展

王万良[1],张兆娟[1],高 楠[1],赵燕伟[2]

(1.浙江工业大学 计算机科学与技术学院,浙江 杭州 310023;
2.浙江工业大学 特种装备制造与先进加工技术教育部重点实验室,浙江 杭州 310014)

摘 要:人工智能、大数据、云计算、物联网等信息技术为推动集成制造快速发展提供了关键技术手段。近年来,采用人工智能技术进行大数据分析取得了突破性进展。系统总结了基于人工智能技术的大数据分析方法的最新研究进展。从大数据的聚类、关联分析、分类和预测 4 个主要的数据挖掘任务出发,分析了大数据环境下机器学习的研究现状;针对深度学习这一热点,总结了基于 MapReduce、Spark 的分布式深度学习实现,以及面向大数据分析的深度学习算法改进相关研究;从群智能、进化算法两方面梳理了基于计算智能的大数据分析相关研究;针对大数据平台,特别对大数据分析和深度学习集成框架进行了归纳,介绍了大数据机器学习系统和算法库;分析了大数据分析中人工智能技术面临的主要挑战,并提出了进一步的研究方向。

关键词:大数据;人工智能;机器学习;深度学习;计算智能

中图分类号:TP301 **文献标识码**:A

Progress of big data analytics methods based on artificial intelligence technology

WANG Wanliang[1], *ZHANG Zhaojuan*[1], *GAO Nan*[1], *ZHAO Yanwei*[2]

(1. College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China;
2. Key laboratory of Special Purpose Equipment and Advanced Manufacturing Technology, Ministry of Education, Zhejiang University of Technology, Hangzhou 310014, China)

Abstract: Artificial intelligence, big data, cloud computing, Internet of things and other information technologies promote the development of integrated manufacturing. Remarkable achievements were achieved in the methods of big data analytics with artificial intelligence technology. The latest research progress of big data analytics methods based on artificial intelligence was summarized comprehensively. A summary of research on machine learning was respectively introduced at first, including big data clustering, correlation analysis, classification and prediction. For the deep learning, a hotspot of research in machine learning, distributed deep learning models based on MapReduce/Spark and other improved deep learning algorithms for big data were discussed especially. The big data analytics based on computational intelligence were discussed from swarm intelligence and evolutionary algorithms two aspects. Furthermore, the engineering implementation of distributed computation platforms for big data were described, including the integrated frameworks for the distributive deep learning, big data machine learning systems and algorithms library. The challenges and the possible research directions of artificial intelligence technologies for big data analytics were put forward.

Keywords: big data; artificial intelligence; machine learning; deep learning; computational intelligence

收稿日期:2017-09-15;**修订日期**:2018-08-01。Received 15 Sep. 2017; accepted 01 Aug. 2018.

基金项目:国家自然科学基金资助项目(61873240,61572438,61702456)。**Foundation items**: Project supported by the National Natural Science Foundation, China(No. 61873240,61572438,61702456).

1 问题的提出

近年来,“两化”融合的不断深入促进了以云计算、物联网、大数据为代表的新一代信息技术与现代制造业等的融合创新。工业物联网、信息物理网络加速了计算机集成制造工厂智能化,实现了企业互联、生产设备之间互联、设备和产品互联、虚拟和现实互联,使得人、机、物信息交互源源不断地产生出工业大数据,进入了“工业大数据”时代。如何通过产品全生命周期工业大数据的采集,获取在线设备全过程、全时段的异构、同步海量数据,并结合人员、环境信息,从全景式大数据分析角度,提取关键性的质量影响因子,并提供更系统有效的质量管控技术;如何通过大数据分析,从产品全生命周期、最终产品的全程供应链与实际应用环境的质量感知出发,针对产品在不同环节和不同应用领域构建多维、精准的质量评价标准并提供相关分析、评价方法与技术,实现产品创新,提升产品质量和品牌,改变中国制造“大而不强”的现状,是中国未来计算机集成制造向更高层次发展的重要需求。

大数据是人类发展的重要经济资产,隐含着很多在小数据时不具备的深度知识和价值,对大数据的智能分析与挖掘将带来可观的经济效益。面对海量、复杂的数据,Google 首席经济学家 Hal Varian[1] 指出:“数据已经变得无处不在,而数据创造的真正价值在于我们是否能够提供数据分析这种增值服务”。大数据价值链中关键的就是对数据的分析,其目标是发现数据的规律,挖掘数据中隐藏的信息,从而辅助制定决策等[2]。智能制造领域是当前人工智能技术重点研究和应用的方向,伴随着海量数据资源和计算能力的提升,将会加速大数据时代人工智能技术在智能制造领域的应用发展[3]。人工智能正在对生产流程、生产模式和供应链体系等制造过程产生巨大影响。工业大数据能够降低集成制造成本、缩短生产周期、提升生产效率、优化生产过程与决策,在智能制造研发设计、生产制造、经营管理、市场销售、售后服务等环节,分析感知用户需求、提升产品附加值,实现智能工厂。总之,基于大数据的智能制造是对制造设备本身以及产品制造过程中产生的大数据进行系统分析,对工业大数据的分析应用将渗透到整个智能制造业价值链,进而为整个制造业的转型升级带来巨大推进力。

人工智能技术是进行大数据分析的一种重要方法,机器学习、计算智能都属于人工智能技术中重要的分析方法,其中深度学习具有卓越的性能,是目前最重要的机器学习方法。大数据平台、人工智能和大数据应用领域之间的关系如图 1 所示。本文主要从大数据分析领域的研究主流——人工智能技术这个维度出发,以当前大数据平台和大数据分析方法为主要研究内容。大数据平台的核心主要包括数据分布式存储和分布式计算两部分。针对结构化、半结构化、非结构化 3 类数据,可基于分布式文件存储系统(Hadoop Distributed File System, HDFS)、分布式数据库 HBase 等实现数据的分布式存储。分布式计算环境可由 Hadoop 集群或虚拟的云资源等支撑,MapReduce 主要用于离线任务的分布式计算、Spark 用于实时在线任务的分布式计算。基于底层的大数据平台,可开展大数据环境下基于机器学习的数据挖掘、基于计算智能的分布式优化等大数据分析任务,并应用于工业智能制造、智能交通、医疗等不同领域。

2 基于机器学习的大数据分析

机器学习是人工智能最重要的分支,也是大数据分析中最重要的方法。国内外已有不少学者对机器学习方法应用于众多领域的大数据分析进行了研究[4-6]。从大数据挖掘主要任务出发,基于机器学习的大数据分析分为 4 个方面:①大数据聚类;②大数据关联分析;③大数据分类;④大数据预测。

2.1 大数据聚类

大数据往往是跨学科、跨领域、跨媒体的,传统聚类算法难以直接应用于大数据聚类,因此,大数据的聚类受到越来越多的关注。

MapReduce 是主流的分布式计算框架之一。基于 MapReduce 实现传统聚类算法的并行运算是大数据分析的一类重要方法,其主要思想为首先针对大规模数据进行数据分块简化处理,再将处理结果合并,即基于 MapReduce 分布式计算框架实现了数据的并行化。Zhao 等[7] 基于 Hadoop 平台实现了经典的 K-means 聚类算法,整个过程主要分成 Map、Combine 和 Reduce 三段。Gao 等[8] 用 MapReduce 编程框架实现了自底向上的凝聚式层次聚类分析(Agglomerative Hierarchical Clustering, AHC)算法,提升了文本聚类时的准确率和召回率。He 等[9] 基于 MapReduce 实现了具有噪声的基于密度的聚类方法(Density-Based Spatial Clustering of Applications with Noise, DBSCAN),主要包括数据预处理、局部 DBSCAN、获取需

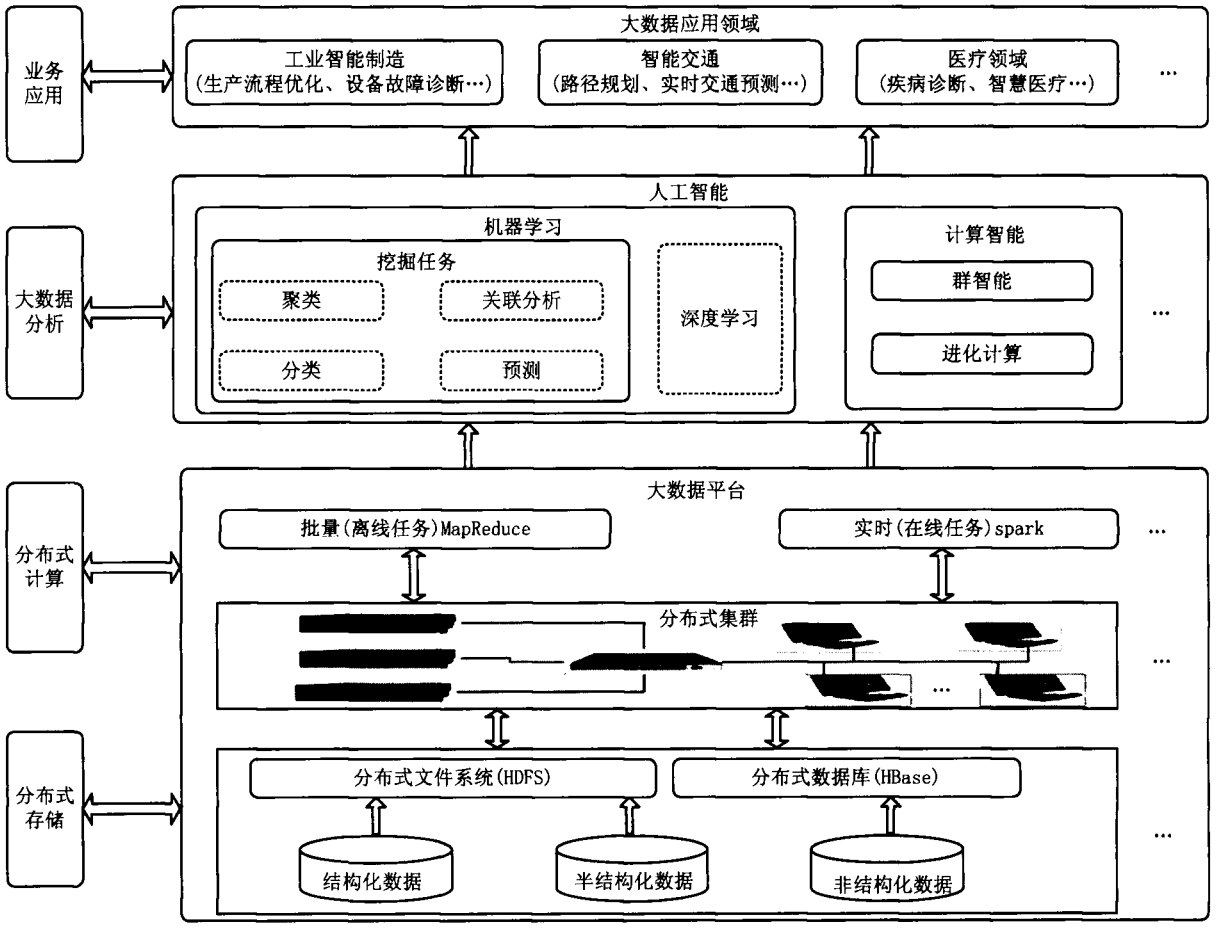

图1　基于人工智能技术的大数据分析主要框架

要合并的集群、全局进行集群处理 4 个阶段,并将其应用于轨迹聚类。

基于 MapReduce 的大数据聚类算法的分布式实现,能够提升算法的聚类效果,并降低计算复杂度。Yan 等[10]提出一种基于 MapReduce 并行幂迭代聚类(Parallel Power Iteration Clustering, P_PIC)方法,使用并行策略增加数据的可伸缩性,在并行策略和实现过程中以计算和通信时间最小化为目标,降低对硬件的需求。Zhao 等[11]基于 MapReduce 改进最大期望算法,用于并行训练概率潜在语义分析模型,从而使每台计算机内存只需加载部分数据,减少数据读取时间与存储容量。Kim 等[12]基于 MapReduce 改进密度聚类(Density-Based Clustering, DBCURE)算法,提出 DBCURE-MR 算法,能够并行发现多个聚类,提高查找聚类效率。Hu 等[13]提出一种快速相似度测量轨迹方法,基于 MapReduce 将轨迹聚类分成 Map、Shuffle 和 Reduce 三个阶段,结合改进的动态时间规整算法进行聚类。Bu 等[14]结合快速搜索聚类(Cluster Fast Search, CFS)算法和下降深度学习模型,提出一种适用于异构大数据聚类的高阶 CFS 算法,首先通过自适应下降深度学习模型提取数据特征,然后运用特征张量模型发现异构数据间的关联。

K-means 算法是一种最典型的聚类算法,应用非常广泛,许多研究着重于提升大数据处理的速度和性能。Liao 等[15]提出 K-means 集群优化算法,通过减少迭代次数以及提高迭代速度来提高算法效率。Cui 等[16]改进了 K-means 算法,提出一种基于 MapReduce 的并行聚类模型,克服了 MapReduce 不适合迭代计算的不足,以消除 K-means 算法对迭代的依赖,从而提升性能。Akthar 等[17]针对 K-means 算法中随机初始中心引起的不稳定性,基于数据维数密度来优化初始中心的选择,提出一种改进 K-Means 聚类算法,能够提高聚类效果,减少时间。Xia 等[18]针对出租车轨迹大数据,提出基于 MapReduce 优化 K-means 聚类方法,将轨迹数据处理流程分成 3 个阶段,采用平方误差准则评判聚类效果,该改进算法在效率和准确率上都优于传统 K-means 和并行两阶段 K-means 算法[19]。

由上述分析可以看出,传统聚类算法在大数据

时代存在数据量大、复杂度高等难题，采用并行实现或改进现有的聚类算法，是当前大数据聚类算法研究的重要方向。

2.2 大数据关联分析

关联分析又称关联挖掘，即在各种数据中查找存在于项目集合或对象集合之间的频繁模式、关联、相关性或因果结构，是大数据挖掘的主要任务之一。当前，应用较多的关联分析算法主要包括 Apriori 关联规则挖掘[20]（广度优先）和频繁模式增长（Frequent Pattern Growth，FP-Growth）算法关联规则挖掘[21]（深度优先）。但传统的串行算法多次扫描数据库，I/O 负载过大，而且效率低，随着数据规模的增大，计算能力和存储容量成为关联挖掘的阻碍。为克服传统单机环境下的挖掘瓶颈，针对大数据进行关联分析，可采用 MapReduce 或 Spark 分布式计算框架。本节对这两种算法的分布式和并行化处理以及一些应用进行了梳理。

2.2.1 Apriori 关联规则挖掘

Apriori 算法先设定一个支持度阈值，再依据该阈值筛选相关事务数据集，进而发现所有的频繁项集，并得到相应的关联规则。Apriori 算法主要有两方面不足：①时间开销大；②会产生大量的候选频繁集，从而在广度和深度上的适应性不好。当处理大数据时，这两个缺点会更加突出。

将 Apriori 算法进行并行化是提升计算效率的重要策略。Li 等[22]提出了基于 MapReduce 的并行 Apriori 算法，将产生候选项集的过程进行并行化，从而提升了 Apriori 算法的运行效率。Ezhilvathani 等[23]和 Oruganti 等[24]基于 MapReduce 实现了 Apriori 算法的并行计算。HAO 等[25]提出基于 MapReduce 下改进的 Apriori 算法，算法效率高、加速比好，但对集群之间的通信具有较高要求。Xiao 等[26]提出一种并行 Apriori 算法，但集群之间的通信使得算法实现很困难。Qiu 等[27]提出了基于 Spark 的并行频繁模式挖掘算法（Yet Another Frequent Itemset Mining，YAFIM）。She 等[28]提出基于 Hadoop 的 Apriori 约简并行改进算法，有效约简了事务数据库，使用哈希树减少了计数时间，提高了算法效率。

改进 Apriori 算法并行化聚类挖掘性能是近年来的重要研究领域。Zhou 等[29]基于 MapReduce 对 Apriori 算法中的连接步进行并行化处理，但是对候选项集的剪枝步并没有进行并行化处理。米允龙等[30]将分治思想与否定粗糙关联规则结合，提出基于 MapReduce 的否定粗糙关联规则算法，能够快速、简便地挖掘出频繁项集后的否定关系。Padillo 等[31]提出基于 MapReduce 框架的 Apriori-Inverse 算法（Apriori-Inverse using MapReduce，Apriori-Inverse-MR），采用支持度低于最小阈值、高于最大阈值、置信度高于阈值的策略，实现了大数据中罕见关联规则的挖掘。Feng 等[32]基于 MapReduce 和 HBase 提出了改进的 Apriori 算法（MH-Apriori 算法），利用 HBase 的时间戳从而避免生成大量的键/值对，提高了效率。Singh 等[33]将 Apriori 算法基于 MapReduce 的实现方式进行分类，主要分为 Map 和 Reduce 算法。

2.2.2 FP-Growth 关联规则挖掘

FP-Growth 算法的基本思想是分而治之（称为分治策略），该思想在处理大数据问题时非常有效，尤其是随着分布式和并行计算的飞速发展，其重要性日益突出。通过构建频繁模式树（FP-tree）并挖掘 FP-tree 上的频繁模式就可以得到数据集内所蕴含的关联规则，算法只需扫描两次数据集即可。与 Apriori 算法相比，FP-Growth 算法避免了频繁模式搜索过程中大量候选项集的产生，但在面对海量数据时，对 FP-Growth 算法进行分布式并行化处理十分必要。

改进 FP-Growth 算法的分布式和并行化处理过程，能够缩短计算时间。Zhou 等[34]基于 MapReduce 对 FP-Growth 算法进行负载均衡改进，但由于并行 FP-Growth 算法在挖掘频繁模式的步骤中，仅保留了支持度较高的 K 个频繁模式，致使其挖掘结果不完备。Xiao 等[35]提出一种基于 MapReduce 框架下的并行 SON 频繁项挖掘算法。王洁等[36]通过键值存储策略优化了 FP-Growth 算法的计数和分组两个过程，并将其并行运行。Wang 等[37]提出一种基于 MapReduce 的 FP-Growth 改进算法（Frequent Itemsets Mining MapReduce，FIMMR），首先将每个独立数据块作为候选项集，挖掘局部频繁项集来确定全局频繁项集。刘智勇等[38]提出了基于 Spark 的并行 FP-growth 算法，通过分组策略将大数据集切分成小数据集，然后分别在小数据集上并行执行 FP-growth 算法得到频繁项集。

2.2.3 关联规则挖掘的应用

关联规则挖掘具有广泛的应用领域，如智能交通、数值分析、疾病诊断、日志分析等。

针对出租车运行轨迹问题,Xia 等[39]提出基于 MapReduce 的面向海量小文件处理策略的并行频繁模式增长算法,对车辆运行的时空特征进行关联分析,通过频繁项集的并行挖掘,提高产生关联规则的效率,具有更好的加速比性能和更高的挖掘效率。针对动车组故障诊断问题,Bin 等[40]提出一种基于 MapReduce 改进的并行 FP-Growth 算法,通过局部频繁模式树代替全局频繁模式树的策略,能够避免因全局频繁模式树过大而造成的计算速度过慢的问题。Bechini 等[41]提出一种基于 MapReduce 的分布式关联规则分类策略,在执行分类关联规则(Classification Association Rules, CARs)的挖掘时,利用改进的 FP-Growth 算法对分布式规则进行修剪,再将存活下来的 CARs 集合用于对未标记的模式进行分类。

窦蒙等[42]针对过程挖掘领域的事件日志大数据,提出基于案例拆分的改进日志转化算法,并基于 MapReduce 对海量事件日志进行分布式处理。Sun 等[43]提出基于 MapReduce 改进的概率后缀树(Probabilistic Suffix Tree, uPST+MR)算法,实现不确定序列数据时隐藏模式的挖掘。Agbehadji 等[44]将狼搜索算法用于数值数据关联规则的挖掘中,能够在不陷入局部最优的情况下,从数值数据中找到最佳规则。Nguyen 等[45]提出一种基于晶格结构和两组对象标识符之间的差异来挖掘具有项目集约束的分类关联规则的方法,从而有效分类 HIV 感染的高风险人群。Lee 等[46]利用关联规则研究了意大利卡利亚里市往返车辆共享行为特征,探索了往返车共享关键因素之间的综合关系,从而最大限度地提高了私家车的使用效率,进一步缓解了城市拥堵。Weng 等[47]提出一种基于关联规则的方法来分析工作区碰撞事故伤亡事件的特征和影响因素,能够提供更多可理解的结果。

综上所述,现有关联挖掘算法的研究主要集中于对已有算法进行分布式和并行化处理,不同的分而治之与并行处理策略会影响到集群的负载均衡与计算效率,但目前的研究主要基于 MapReduce 平台。Spark 作为一个基于内存计算的大数据平台,计算速度比 MapReduce 平台快,进一步可考虑基于 Spark 进行关联分析算法的分布式和并行化处理,以及针对大数据的特点,设计新的分布式和并行化关联规则挖掘算法。

2.3　大数据分类

大数据分类是大数据挖掘中的一种重要手段。大数据的分类问题普遍存在,应用在各行各业,如网络入侵检测、医疗诊断等。

Del 等[48]基于 MapReduce 实现了随机森林方法,用于非均衡数据的分类。选择最佳分裂属性是整个决策树生成中最耗费计算资源的部分。决策树算法的并行化实现中,关键的就是利用 MapReduce 框架对选择最佳分裂属性这一过程进行加速。SINGH 等[49]使用 Mahout 的并行处理能力来构建基于决策树模型的随机森林,应用于点对点僵尸网络的实时检测。

López 等[50]提出代价敏感的惩罚确定因子方法,在计算规则权重时考虑误分类的代价,基于 MapReduce 和代价敏感学习策略,提高了分类准确性,降低了时间成本。Huang 等[51]对极限学习机(Extreme Learning Machines, ELM)在大数据分类的研究进展进行了分析,由于 ELM 具备良好的泛化性能,可扩展到代表性学习、功能选择和其他学习上。Kamal 等[52]将 K 近邻分类器和 MapReduce 相结合,用于 9 000 万对 DNA 的分类。针对辅助医疗诊断肿瘤问题,Kumar 等[53]将基于 MapReduce 的 K 近邻分类器用于对微阵列基因进行分类,从而分析是否携带癌症基因。

Fernández 等[54]评估了 17 个派系的 179 个分类器,涉及判别分析、贝叶斯、神经网络等经典方法。使用来自加州大学欧文分校(UCI)数据库的 121 个数据集来研究不同分类器的作用,表明 R 语言实现的随机森林(RF),以及 C 语言实现的支持向量机的分类效果较好。Hafez 等[55]基于 Spark 平台,探讨了不同类型的机器学习算法适用于哪种应用范围。针对市场在线营销、包装和统计、安全 3 类数据集,主要依据准确性和训练时间来评判分类和回归效果,其中决策树算法适用于市场营销和安全数据集的分类,逻辑回归算法适用于包装和统计类数据的分析。

大数据环境下,由单一数据逐渐过渡到分布式数据集,各种分类算法都面临着大数据环境的挑战,因此传统分类方法很难直接运用到大数据环境下。基于机器学习的大数据分类是当前的研究热点,怎样结合大数据平台,将机器学习应用于不同领域的分类是一个极具挑战的难题。

2.4 大数据预测

大数据预测是大数据研究的核心内容之一，其可以应用于很多行业，如价格预测、网络入侵检测、化学元素分析、医学、电力负荷预测、智能制造车间运行状态预测、企业绩效分析等。

基于机器学习的大数据预测应用十分广泛。Ruta 等[56]将机器学习算法用于金融领域，针对流式大数据，基于多个市场间的相关性和市场结构的差异性，建立了一个可扩展的交易模型，运用逻辑回归方法进行实时价格预测。Suthaharan 等[57]运用大数据预测网络入侵，提出结合 HDFS、云计算、几何特征学习等技术，运用支持向量机对网络中可能出现的入侵攻击进行预测的算法。Ramakrishnan 等[58]将机器学习运用到量子化学领域中，提出 Δ-ML 模型，即用机器学习方法纠正传统量子方法中代价近似这一环节，选取较大的分子集进行训练，从而高度准确地预测焓、自由能和熵。

采用基于机器学习的大数据预测可以实现精准医疗。Google 构造出一个流感预测指数模型，成功“预测”了流感病人的就诊人数[59]。Bibault 等[60]采用支持向量机方法对放射肿瘤学模型进行综合分析预测，提升了放射治疗系统的安全性和治疗效果。Zhu 等[61]首次提出一种全尺寸、无标注、基于病理图片的病人生存有效预测方法(Whole Slide Histopathological Images Survival Analysis, WSISA)，在肺癌和脑癌两类癌症的 3 个不同数据库的性能均超出基于小块的图像方法，实现了基于大数据分析的精准个性化医疗。

采用机器学习可以对电网负荷进行预测。Simmhan 等[62]基于动态需求响应(Dynamic Demand Response, D2R)云平台，首先通过语义信息集成机制获取动态数据，将可扩展的回归树模型用于训练大量历史数据，对 Web 和移动 App 门户网站的当前和历史电能消耗进行预测，从而缓解高峰负荷并实现校园内的智能用电需求管理。

通过大数据处理与分析，能够深入掌握数据间的复杂关联关系，对智能工厂车间制造过程的性能变化进行预测与调控，实现智能制造。杨俊刚等[63]提出面向半导体制造的大数据分析平台，实现了工期预测、晶圆良率预测等。张洁等[64]针对智能车间制造数据呈现的大数据特性，研究了海量高维多源异构制造数据分析方法和大数据驱动的车间运行状态预测、决策方法等。吕佑龙等[65]提出基于大数据的智慧工厂技术框架，探讨了大数据驱动的制造过程动态优化关键技术。姚锡凡等[66]引入主动计算和大数据分析等技术，提出一种大数据驱动的新型制造模式—主动制造(proactive manufacturing)。朱雪初等[67]提出了基于工业大数据的晶圆加工周期预测方法框架。Wamba 等[68]提出一种工业大数据预测模型，有效提升了企业绩效。

大数据预测目前存在两个主要困难：①在预测时，快速获得一个大概的轮廓和发展趋势比获得精确的结果重要，但在需要根据大数据进行个性化决策时，精确性则变得非常重要。在进一步研究的新方法中，需要在效率和精确性之间找到平衡点。②在大数据中，存在的有价值信息与数据规模的扩大并不是成比例增长，从而导致获取有价值信息的难度加大。例如，在连续监控的视频中有价值的数据可能只有几秒。怎样在大数据中找到这些有价值的信息是提高大数据分析方法性能的关键。

综上所述，机器学习是人工智能的一个核心研究领域，基于机器学习的大数据分析方法是当前发展最为迅速的方法。结合 MapReduce 和 Hadoop，将机器学习应用于大数据聚类、关联分析、分类和预测各个方面，取得了突破性进展。但针对基于机器学习的大数据分析，怎样设置 Map 和 Reduce 函数合适的输入输出键值对是亟待研究的重要问题。

3 基于深度学习的大数据分析

自 2006 年，加拿大多伦多大学教授 Geoffrey Hinton[69]在《Science》上提出深度学习以来，深度学习便成为机器学习最重要的研究领域。深度学习是最重要的机器学习方法之一，广泛应用于图像、语音、自然语言处理领域，笔者 2016 年出版的《人工智能及其应用》(第三版)教材中比较系统地介绍了深度学习的主要内容[3]。

深度学习的训练是一个计算密集型的任务，模型训练过程中确定各个隐层的权值、阈值等参数都需要经过大量的迭代计算。对于中等规模，即具有几个隐层和每层包含数百个隐层节点的深层网络，学习可能会消耗几天甚至几周[70]。当数据规模扩大时，整个模型的训练所需时间大大增加。由于深度学习训练的数据量非常庞大、整个训练过程耗时，而 MapReduce、Spark 平台支持分布式计算、HDFS 能够实现分布式文件存储。因此，将深度学习算法和大数据平台结合，通过分布式来降低深度学习的

训练时间成本引起了学者的广泛关注。下面对基于 MapReduce 的深度学习分布式实现、基于 Spark 的深度学习分布式实现和面向大数据分析的深度学习算法的改进进行介绍。

3.1　基于 MapReduce 的深度学习分布式实现

由于计算代价和时间成本快速增长,而 MapReduce 分布式计算框架更易于处理大量数据,因此,基于 MapReduce 的深度学习分布式实现变得越来越重要。

目前关于基于 MapReduce 的分布式深度学习的研究较少,主要集中在深度学习一些基础模型方面。为克服受限玻尔兹曼机(Restricted Boltzmann Machine, RBM)预训练非常耗时的不足,Zhang 等[71]提出一个针对 RBM 和反向传播(Back Propagation, BP)的分布式学习范式,基于 MRjob 编程框架,首次基于 MapReduce 框架实现了深度置信网络(Deep Belief Nets, DBNs),采用批量更新和数据并行的方式来训练深度神经网络,显著缩短了 DBNs 的训练时间。Zhang 等[70]基于 MapReduce 实现了 RBM 的分布式学习,其分布式学习策略有效解决了深度学习的可拓展问题。

王德文等[72]开发了基于 Hadoop 的大数据并行负荷预测系统,利用分布式计算框架 MapReduce 对电力用户侧的大数据进行分析,缩短了负荷预测时间。针对采用 BP 算法训练的全连接多层神经网络,Zhang 等[73]提出一种针对 Map 阶段的有效映射机制,并基于 MapReduce 实现了 BP 算法的分布式,从而以更快的速度收敛。Cao 等[74]将并行粒子群算法用于优化 BP 神经网络的初始权重和阈值,并基于 MapReduce 实现了 BP 神经网络的并行处理,从而解决了 BP 神经网络在处理大数据时存在的硬件和通信开销问题。毛国君等[75]基于 MapReduce 设计流式大数据的分布式分类模型和挖掘算子,有效减少了网络节点通信代价。

3.2　基于 Spark 的深度学习分布式实现

由于 Spark 分布式计算平台是基于内存进行计算的,同 MapReduce 相比更适合迭代型计算任务,不少学者对深度学习模型在 Spark 上的分布式实现展开了研究。何洁月等[76]首先建立了基于实值的状态玻尔兹曼机(Real-valued Conditional Restricted Boltzmann Machine, R_CRBM)模型,并针对大数据环境下 R_CRBM 模型训练耗时的不足,提出基于 Spark 内存计算平台的深度学习并行化方案,加速深度学习训练过程。Li 等[77]结合大数据分布式平台 Spark 实现了卷积神经网络,并通过批归一化技术和多交叉测试来修正卷积神经网络,提升了 ImageNet 数据集的分类性能并缩短了时间。

通过深度神经网络构建大数据分析模型,能够揭示大数据中隐藏的丰富而复杂的信息,对未来做出精准预测。杨佳驹等[78]提出一种改进的 DBN 模型,采用信息熵理论来确定其隐层的神经元个数,基于 Spark 平台的弹性数据集(Resilient Distributed Datasets, RDD)分区特性将电力大数据进行分块,实现电力系统的短期负荷预测。

针对移动大数据,Alsheikh 等[79]基于深度学习在 Spark 平台上进行分析。在 Spark 上的深度学习模型分布式实现主要包含移动大数据的分块、深度学习模型并行化两个步骤,整个训练过程需进行多次迭代计算,直至收敛。整个训练过程如下:各个工作节点独立训练部分深度模型,并将学习的计算参数传递给主节点,然后主节点对各个从节点学习的模型参数进行平均,重建主控深度模型。在 Spark 上的深度学习模型分布式实现,加快了深度学习模型的学习速度、缩短了训练时间,也提高了移动大数据行为识别的准确性、降低了错误率。为了对人类移动轨迹大数据进行预测,Ouyang 等[80]提出深度结构模型"DeepSpace",并结合连续的移动数据流,开发了香草卷积神经网络(Vanilla Convolutional Neural Network, VCNN)在线学习系统,采用中国东南部城市移动蜂窝网络的数据记录进行了测试,能够较精准地预测人类轨迹,有助于对人类移动模式进行研究。

随着数据规模逐渐增大,深度学习模型可扩展性的重要性也日益突显。Yan 等[81]针对非均衡数据集中类分布存在倾斜、且分类器倾向于大多数类的不足,首次将引导抽样(bootstrapping)引入卷积神经网络中,拓展了深度学习模型。首先,从训练的神经网络中提取特征,再将该特征运用到另一个全连接神经网络中,提高了卷积神经网络分类的准确性。针对卷积神经网络耗时、不适合流数据的不足,基于 Spark 分布式框架和深度学习工具 Deeplearning4j 进行神经网络训练的分布式计算,缩短了训练时间。改进的深度学习模型对多媒体数据的低层特征具有较好的效果,且具有较好的灵活性和扩展性。

3.3　面向大数据分析的深度学习算法的改进

基于 MapReduce、Spark 分布式计算平台能够

降低深度学习算法的训练时间成本，但面向大数据分析，提高训练样本的质量、加速深度网络的训练过程、优化深度学习模型的性能等仍是深度学习算法的关键问题。

实际应用时，常常存在样本数量缺乏、样本质量不足等问题。为实现类脑的高级智能，怎样提升大数据训练样本的数量和质量引起了学者的关注。Yang 等[82]提出固定模型重用策略，即用深层模型的学习能力从固定模型中获取有用的判别信息来减少对训练样本的需求。Jie 等[83]提出一种依靠检测器自身不断改进训练样本质量从而增强检测器性能的方法，解决了弱监督目标检测时训练样本质量较低的问题。

通过对深度学习模型自身进行改进，也能够加速深度网络的训练进程。Ioffe 等[84]通过对输入层进行批归一化(normalizing layer inputs)从而解决深度神经网络训练过程中内部协变量的转变(internal covariate shift)，加速了深度网络的训练。He 等[85]提出一种残差学习框架，表示为与输入层相关的学习残差函数(learning residual functions)，从而简化了深度神经网络的训练，这种残差网络更易于优化，并能通过增加层数获得更高的准确度。进一步，Kranjc 等[86]开发了一个分布式计算平台“群流”(ClowdFlows)，用于提高构建和执行流数据挖掘的计算速度。

多种模型的混合以及对模型自身的改进有助于提升深度学习算法的性能。Mnih 等[87]将强化学习中的 Q-学习与卷积神经网络混合，提出深度 Q-网络，在视频游戏中能够达到与专业人类玩家相当的水平。Xue 等[88]提出具有跟踪的概率运动模型学习分类器，通过引入概率运动模型且最大化后验概率方法，提升深度卷积神经网络用于视频中行人的监测和跟踪效果。Zhao 等[89]提出了无分布一次通过学习(Distribution-Free One-pass Learning, DFOP)方法，用以解决在线学习数据增长带来的基础分布发生变化的问题，无需保留处理过的数据就能取得显著的分类效果。Ke 等[90]改进侧输出残差网络从而提升了多尺度特征融合性能。Zhou 等[91]设计了具有主动旋转能力的向量场滤波器(Active Rotating Filter, ARF)，显著提升了深度网络特征对旋转的泛化性能。Finn 等[92]提出一种与模型无关的学习算法(Model-Agnostic Meta-Learning, MAML)，它能匹配任何使用梯度下降法训练的模型，在少量样本、较少迭代次数下就能具有很好的泛化性能。

综上所述，深度学习有助于提升对大数据挖掘的精度和深度，是对大数据进行分析的一种重要分析方法。伴随着大数据时代的来临，深度学习取得了重大突破，拓展了人工智能所能解决问题的边界，分布式深度学习将会越来越重要。针对大数据分析，怎样改进深度学习算法从而提升其性能值得进一步深入研究。

4 基于计算智能的大数据分析

计算智能是人工智能的一个重要分支，由于启发式和随机特点，计算智能非常适合解决大规模优化问题。传统的优化算法多采用集中式的设计思想，主要考虑算法的收敛性和收敛速度。当所要解决的问题非常复杂，或者规模很大时，尤其是大数据所面临的问题，传统的集中式优化算法无法处理或者计算非常耗时。数据规模和复杂性的日益增加给传统的计算智能算法带来了新的挑战，因此，需要研究分布式优化算法来解决大数据优化所面临的问题。下面主要从群智能和进化计算两个方面，总结基于计算智能的大数据分析方法的相关研究。

4.1 基于群智能的大数据分析

群智能是面向海量、高维、动态特性大数据分析的一种重要方法[93]。基于群智能算法的分布式实现是大数据分布式优化中的一个分支，由于基于分布式计算环境能够加速优化算法搜索的过程，在大数据优化问题中占据重要地位。基于群智能的大数据分析方法目前主要是基于粒子群优化算法的大数据分析方法，此外还有基于蚁群算法、布谷鸟算法、萤火虫算法、猫群算法等大数据分析方法。

4.1.1 基于粒子群优化的大数据分析

粒子群算法是最重要的一种群智能算法，如何进行分布式实现是当前的重要研究领域。McNabb 等[94]首次基于 MapReduce 对粒子群算法进行分布式实现，在 Map、Reduce 阶段怎样设计合适的键值对较为关键。基于 MapReduce 进行粒子群算法的分布式实现，对于大数据分布式优化具有重要的意义。

面向大数据的分析一般是非常困难的，采用分而治之的方式是处理高维数据集上粒子群优化问题的一个有效策略。Liang 等[95]提出一种拆分与融合策略。Li 等[96]使用动态的随机分组策略，将高维空

间划分为大小可变的低维子空间,提出一种适用于大规模高维数据空间优化问题的协同演化粒子群算法。徐宗本等[97]提出一种数据拆分与融合策略,就如何划分样本子集规模、如何保持子集之间的信息传递、如何设计各子集结果的融合等问题进行了改进。

基于目标函数的聚类可视为优化问题,从而采用群智能算法进行求解。Govindarajan 等[98]根据效率、准确性和错误计数 3 个因素,基于 MapReduce 运用粒子群算法对学生数据进行分布式并行聚类。Gupta 等[99]从 HDFS 的数据存储、群智能的并行计算、任务调度出发,对群智能算法在 MapReduce 等大数据平台中的运用进行分析,实现了粒子群优化算法对 twitter 数据集的聚类。Wang 等[100]将大数据挖掘中的特征选择视为大数据驱动的组合优化问题,并将其用于高维数据搜索和减轻生物信息大数据的处理负荷中。

Cheng 等[101]分析了群智能算法大数据工程中应用的可能性。由于大数据具有"4V"特点(体量大、速度快、模态多、价值大密度低),对应于群智能算法,分别表示为大规模、高维、动态、噪声/不确定/代理。以大数据驱动的港口物流和旅行商问题作为实例,运用粒子群算法进行物流路径规划和问题求解,都取得了较好的效果。Fong 等[102]提出通过加速的粒子群优化算法进行群搜索,解决了高维流式大数据搜索速度较慢的不足,提高了分析精度。

采取维度分布式策略能够加速对决策空间解的搜索,基于 MapReduce、Spark 分布式计算框架能够缩短协同粒子群优化算法的搜索进程[103-104]。进一步,针对量子智能等较新算法的出现,Li 等[105],Ding 等[106]基于 MapReduce 探讨了基于吸引子、量子旋转门这两种不同编码和再生方式下的量子粒子群优化算法的分布式实现策略。

4.1.2　基于其他群智能算法的大数据分析

类似于粒子群算法,可以采用其他群智能算法进行大数据分析。目前主要集中于群智能算法在 MapReduce 编程框架上的分布式实现,用于提高算法运行效率、加快求解速度。当数据量具有一定规模时,基于 MapReduce 的群智能算法在收敛速度上优于非分布式实现。

蚁群算法是一种本质上并行的算法,具有较强的全局搜索能力。由于蚁群优化(Ant Colony Optimization, ACO)收敛速度较慢,且容易陷入停顿状态,Cheng 等[107]提出一种动态正负反馈蚁群优化算法,即采取在内部正反馈、群体之间负反馈的策略,并在使用迭代 MapReduce 模型构建的框架 Haloop 上分布式实现。吴昊等[108]将分治策略和模拟退火算法共同引入蚁群算法中,提出基于 MapReduce 的蚁群算法(MapReduce-based ACO, MRACO),提高了蚁群算法处理大数据的能力。马文龙等[109]针对云制造的动态服务组合优化问题,提出一种改进的蚁群算法,即在蚁群算法中引入最优路径列表和轮盘赌的选择机制。

布谷鸟搜索算法的全局搜索使用 Lévy 飞行,能够以较高的概率发现全局最优。Lin 等[110]基于 MapReduce 改进了布谷鸟搜索(MapReduce Modified Cuckoo Search, MRMCS)算法,同基于 MapReduce 的粒子群优化(MapReduce Particle Swarm Optimization, MRPSO)算法相比,搜索速度提升了 2 到 4 倍。Xu 等[111]进一步改进布谷鸟搜索算法,提出一种基于 MapReduce 的新兴启发式布谷鸟搜索算法,缩短了大数据优化求解的搜索时间且提高了性能。

针对萤火虫优化算法处理大规模数据集耗时、低效的不足,Al-Madi 等[112]提出了基于 MapReduce 的萤火虫群优化聚类并行(Glowworm Swarm Optimization Clustering using MapReduce, MRGSOC)算法,基于大数据的聚类任务则等效于找到多重质心,充分利用了萤火虫群适用于解决多模态问题的优势,聚类效果较好。郑宏升等[113]提出一种改进的萤火虫算法,并用于多服务质量的云调度最优方案搜索中,采用动态优先级算法确定任务顺序,减少了工作流的完成时间。Lin 等[114]提出一种改进的猫群优化(Improved Cat Swarm Optimization, ICSO)算法,通过使用交叉操作和位置更新来产生候选解,从而克服了猫群优化算法计算耗时的不足,加快了寻找最优解时的收敛速度。

群智能算法由于自然的分布特性,非常适合于大数据分布式处理,特别是随着量子粒子群优化算法等一批新的群智能算法的出现,优化性能越来越好,因此,基于群智能算法的大数据分析方法会得到越来越多的研究与应用。综上所述,怎样运用大数据框架来进行分布式计算,或者改进已有优化算法,以便在 MapReduce 上进行分布式实现,从而降低计算求解速度是大数据分布式优化的关键。

4.2 基于进化算法的大数据分析

MapReduce不适合迭代计算，但是进化算法却包含了大量迭代计算，因此如何使MapReduce适用于进化算法的迭代计算是需要解决的关键问题。Jin等[115]通过在MapReduce中增加一个Reduce过程，使得改进的MapReduce能够实现遗传算法的分布式，这是首次基于MapReduce实现遗传算法。

对大数据进行分组是提升算法执行效率的重要手段。Yang等[116]从理论上证明了随机分组的策略可以增加相关变量被划分至同一子分量(subcomponent)的概率，并将其应用于大规模协同进化计算中，这一随机分组策略在高维优化问题中具有明显优势。Omidvar等[117]提出差分分组(differential grouping)的自动分组策略，使得不同分组内变量之间的相互依赖度最小化，克服了简单地采用随机分组策略的不足，使分组更加智能。针对基于多目标演化算法的子集选择算法在处理大规模数据时非常耗时的不足，钱超等[118]提出一种基于分解策略的多目标演化子集选择(Pareto Optimization for Subset Selection, POSS)算法，首先将整个子集空间进行分解，再逐步调用POSS算法进行求解，在分解个数增加的同时，运行时间超线性下降。

多种算法的混合搜索能够提高群智能算法搜索性能。Gheyas等[119]提出一种模拟退火和遗传算法的混合算法，结合遗传算法交叉算子具有较高收敛速度、贪心算法具有局部强搜索能力，以及广义神经网络算法具有较高计算效率等优点，用于解决选择最优化特征子集的NP完全问题。Bacardit等[120]总结了大数据集中基于遗传算法的机器学习的改进策略，主要将其分为软件方法、硬件加速技术、并行计算以及以Hadoop为代表的大数据分布式计算4类。Yuan等[121]提出基于MapReduce的遗传决策树优化算法，具有较高的分类精度和较短的运行时间。

在智能制造领域，生产流程的设计、制造资源分配等本质上都是最优化问题。张影等[122]采用量子多目标进化算法对大数据环境下的云联盟数据资源多服务组合问题进行求解。朱李楠等[123]针对云制造资源的优化组合问题，结合云制造实际工况改进差分进化算法，增添块变异、块交叉和块选择操作，实现最大完工时间最小化。

综上所述，计算智能为大规模复杂优化问题的求解提供了有效手段，是人工智能的重要研究方向，应用计算智能方法进行大数据分析具有巨大的潜力，大数据也给计算智能的发展带来了新的挑战与机遇。进化算法在大数据平台上的分布式实现已有的成果不多，主要原因是优化算法一般都是迭代算法，而MapReduce不适合迭代计算，但对于离线批量处理任务，Hadoop依然具有优势。Spark作为一种分布式计算平台，由于基于内存计算的优势，相比MapReduce更适合迭代型计算任务，进一步可考虑基于Spark平台进行优化算法的分布式实现。但通常情况下很难估算和保证Spark运行迭代算法时需要的内存大小，还可以将Spark与Hadoop相互结合，来提高优化算法的运行效率和处理规模。

5 大数据平台

传统单机平台难以满足大数据的要求，大数据平台为大数据应用的工程实现提供了必要的支撑。自2015年底Yahoo开始将深度学习框架和Spark进行结合，以满足深度学习不断增加的计算复杂度的需求，之后不少主流公司都陆续开源了大数据和分布式集成框架。下面对已有的大数据分析和深度学习集成框架、大数据机器学习系统和算法库进行归纳，供研究人员参考。由于所有项目已开源，在GitHub上能找到对应的源码。

5.1 大数据分析和深度学习集成框架

数据规模的不断扩大直接导致了深度学习模型计算复杂度的增加，面向大数据的深度学习模型难以放入仅有单个图形处理器(Graphics Processing Unit, GPU)的计算机中进行运算，需要部署分布式实现。现有的深度学习框架通常部署在单独的集群环境中，深度学习的流程创建多个程序，从而造成不同独立集群之间大型数据集的传递，增加了系统的复杂性和端到端的学习延迟。但在Spark分布式集群中，每个节点读取本地数据，然后进行合并、排序、Reduce操作，能够避免网络间的数据传输，提高了效率。因此，将深度学习框架和大数据分析平台结合在一起，可以有效降低基于深度学习的大数据分析的复杂性，也避免了多个集群间大量的数据传输降低计算效率的不足。

5.1.1 SparkNet

针对深度学习框架Caffe不支持异构、通信代价较高、训练耗时的不足，Moritz等[124]将Caffe和Spark进行集成，建立了一个基于Spark的深度神经网络架构SparkNet。将训练数据处理和深度网

络训练的融合集成到统一的 Spark 大数据框架中,每次迭代都依次运行数据处理和深度学习等差异较大的模块。源码获取网址:https://github.com/amplab/SparkNet。

李亮等[125]改进了 SparkNet 架构,通过轻量级数据存储将独立运行的 Spark 和 Caffe 以松散耦合的形式进行组合,将数据处理和深度学习训练划分到不同的进程运行,发挥各自引擎的计算能力。此外,消除了基于 Spark 的 Java 虚拟机(Java Virtual Machine, JVM)数据处理和基于 GPU 的 Caffe 深度学习之间的频繁现场切换,并通过轻量级数据存储有效复用 Spark 转化后的训练数据,减少了深度学习执行过程中 Spark 重复进行的数据转化工作,加速训练过程。

5.1.2　DeepSpark

Kim 等[126]将 TensorFlow 集成到 Spark 中,构成了一个大数据分布式深度学习集成框架 DeepSpark。DeepSpark 同时集成了 Caffe 和 TensorFlow 两个深度学习框架,自动地分配负载和参数给 DeepSpark 集群中的 Caffe/Tensorflow 节点,采用平均随机梯度优化算法进行参数更新,实现自适应异构变量训练结果的迭代进化。源码获取网址:https://github.com/deepspark/deepspark。

5.1.3　Deeplearning4j on Spark/Hadoop

Deeplearning4j 是一个兼容 Java、Clojure 和 Scala 语言的分布式深度学习框架,可以与 Hadoop 及 Spark 集成,支持分布式 GPUs 和 CPUs[127]。Deeplearning4j on Spark/Hadoop 支持并行迭代算法架构,可以快速处理大量数据,并整合了 RBM、CNN 等深度学习模型,其中 Yan 等将 Deeplearning4j 和 Spark 集成,实现卷积神经网络训练的分布式计算,有效地缩短了深度学习模型的训练时间[81]。Deeplearning4j 项目源码获取网址:https://github.com/deeplearning4j/。

5.1.4　CaffeOnSpark

Yahoo 将 Caffe 与 Spark 集成,形成了 CaffeOnSpark 分布式深度学习框架。CaffeOnSpark 能够在 GPU 和 CPU 服务器集群上进行分布式深度学习,采用 peer-to-peer 通讯,避免驱动器潜在的扩展瓶颈,具备较强的灵活度。CaffeOnSpark 支持已有的 Caffe 用户无缝集成相关的数据集和配置,也可高度集成增量学习、特征学习、以及机器学习方法,深度学习与传统机器学习的数据处理处于同一个集群中,基于 Spark 能够实现大规模数据的深度学习训练,基于 Hadoop 资源管理器优化深度学习资源的调度。源码获取网址:https://github.com/yahoo/CaffeOnSpark。

5.1.5　TensorFlowOnSpark

2017 年,Yahoo 开源了大数据集群的分布式深度学习框架 TensorFlowOnSpark,支持 TensorFlow 和 Spark、Hadoop 的集成,TensorFlow 与 Spark 集群中的数据兼容,支持模型的并行和数据的并行处理。TensorFlowOnSpark 与 SparkSQL、MLlib 和其他 Spark 库都在一个单独流水线或程序中运行,支持 TensorFlow 进程之间的直接通信,过程到过程的直接通信机制使 TensorFlowOnSpark 上的程序能够在增加的机器上简单进行扩展。TensorFlowOnSpark 利用 TensorFlow 的 File Readers 和 Queue Runners 直接从 HDFS 文件中读取数据,Spark RDD 数据传输到每个 Spark 执行器,随后数据将通过 feed_dict 传入 TensorFlow。源码获取网址:https://github.com/yahoo/TensorFlowOnSpark。

5.1.6　Spark on PADDLE

PADDLE 是百度的分布式深度学习平台,能够支持多 GPU/CPU 的训练,相关的训练算法进行了深度优化,训练效率非常高。该 Spark on PADDLE 大数据深度学习平台支持异构,加入了机器学习决策模块做超参数选择等。训练过程中引入了监控机制和容错机制,在异构平台上使用 YARN 对资源进行分配,从而更有效地实现了资源的统一管理。大数据深度学习平台 Spark on PADDLE 作为一个基础架构,Liu 等[128]将其运用在百度的无人驾驶中。其中一个 Spark 驱动程序用来管理所有的 Spark 节点,且每个节点都托管一个 Spark 执行器和一个 Paddle 训练器。源码获取网址:https://github.com/PaddlePaddle/Paddle。

5.2　大数据机器学习系统和算法库

近年来,基于人工智能技术的大数据分析发展迅速,大数据机器学习系统成为研究热点。美国卡耐基梅隆大学机器学习系 Eric Xing[129]认为开发大数据机器学习系统具有重要意义,提出了大数据机器学习系统设计应主要考虑解决如下 4 个问题:①如何分配机器学习程序;②如何将机器学习计算和通信连接起来;③如何通信;④通信的内容。基于上述问题来设计和开发高性能的大数据机器学习系

统，在机器学习正确性、速度和可编程性之间取得了更好的平衡。国内外已经出现许多大数据机器学习系统，主要有南京大学 PASA 大数据实验室的 Octopus[130]、美国卡耐基梅隆大学的 GraphLab[131] 和 Petuum[132]、新加坡国立大学的 Apache SINGA[133]、百度的 ELF[134] 和腾讯的 Angel[135] 等。

针对主流大数据平台 Hadoop 和 Spark 进行经典机器学习算法的并行化设计，有助于数据分析人员提高效率[136]。目前，在各个大数据平台上的机器学习算法库主要有 MLlib[137]、Mahout[138]、H2O[139]、Fregata[140]、SystemML[141] 和 Scikit-learn[142]。由于不同的大数据平台都提供了相应的大数据机器学习算法库，但实现方式各有差异，优缺点不尽相同。大多数机器学习算法都需进行多次迭代计算，在面对此类问题时，MapReduce 平台执行效率较低。Spark 由于中间数据存放在内存中，能够避免不必要的 I/O 操作且有分布式 RDD 抽象，因此对于迭代运算类的算法而言，效率更高。

综上所述，本章从工程实现角度，对当前大数据平台进行了总结归纳。分布式深度学习变得越来越流行，且迅速发展演化，因此特别对分布式深度学习的分析平台，即引起广泛关注的分布式深度学习集成框架进行了梳理。同时，为了工程实现上能够更高效地运用机器学习进行大数据分析，还对主流的大数据机器学习系统和算法库进行了总结。

6 面临的挑战及进一步的研究方向

6.1 面临的挑战

人工智能技术能够处理大数据面临的许多挑战，也给大数据分析方法带来了许多问题，许多针对小数据上的人工智能方法不能直接应用于大数据分析。从大数据分析的角度来看，大数据主要带来了降低分布式计算的时间成本和提升算法的性能两大挑战。

(1)降低分布式计算的时间成本　大数据的一个主要特点就是算法要实施分布式计算。MapReduce 和 Spark 作为分布式计算平台已被广泛使用。针对深度学习，其中 Google MapReduce 架构设计师 Jeffrey Dean[143] 将其训练过程的并行化归纳为数据并行与模型并行两种策略。卡耐基梅隆大学的 Eric Xing 等[129] 将机器学习的并行策略总结为模型并行和数据并行两种。数据并行是指将数据集进行并行化分割；模型并行是指将模型进行拆分，由训练节点分别持有，通过共同协作完成训练。类似地，针对计算智能算法，Gong 等[144] 将分布式策略归纳为两种：①种群分布，即将种群(或子种群)的个体分布到多个处理器或计算节点；②维度分布，即在数据层面上的并行，主要对搜索空间进行分块并采用协同实现进化。

影响大数据分析算法的时间成本，除了选择怎样的分布式计算平台之外，采取怎样的分布和并行策略非常关键。基于机器学习特别是深度学习、计算智能等基于人工智能技术的大数据分析方法可以进一步精心设计相应的并行、分布式策略来有效降低计算时间成本，如图 2 所示。由于数据样本往往独立同分布，大多数研究都在数据层面上实现算法的并行化和分布式。而实际上，针对机器学习、深度学习算法，怎样实现模型并行却比较复杂，如模型怎样进行拆分、参数怎样进行更新都跟算法本身紧密相关；针对计算智能算法，采取怎样的种群分布模型也会较大程度地影响算法计算时间成本。因此，采取怎样的分布式、并行策略来降低计算时间成本是基于大数据分析方法面临的一个重要挑战。

(2)提升算法的性能　降低分布式计算时间成本并提升算法的性能是大数据分析方法的主要目标，但时间成本和算法性能之间往往难以平衡。怎样提升算法的性能并取得较优的应用效果是大数据分析方法面临的另一个重要挑战。影响算法性能的因素同算法自身密切相关，可进一步对算法进行改进，如优化机器学习、深度学习等算法的重要参数，引入新的进化机制并融合于计算智能算法，以及混合多种算法等都有助于提升算法的性能。

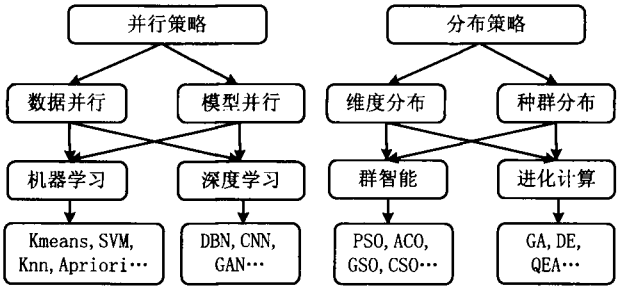

图2　基于并行、分布式策略降低算法时间成本

6.2 进一步的研究方向

虽然人工智能技术是大数据分析的利器，但面临大数据问题时，现有的机器学习、深度学习、计算智能等人工智能分析方法、大数据平台都存在许多不足，难以有效解决大数据的诸多问题。目前进一步研究的主要方向有：

(1)*分布式深度学习算法*　大数据平台以分布式计算和分布式数据存储满足了大数据的处理需求。基于大数据平台 Spark,能够有效缩短深度学习模型的训练时间,可将更多的深度学习模型运用于大数据分析任务中。不断出现的新的深度学习模型应用于大数据分析仍然是当前的主要研究方向。

(2)*设计机器学习模型并行策略*　由于在 Hadoop 和 Spark 分布式集群中的 Map 和 Reduce 阶段结束时,都需要进行参数的更新,哪些参数采取并行更新以及更新的顺序,都会对算法的性能产生影响。如何设计机器学习模型并行时的参数更新策略,是基于机器学习的大数据分析方法中一个亟待解决的关键问题。

(3)*分布式优化算法*　分布式计算在大数据处理中呈现出绝对优势,数据量越大,优势越明显。由于分布式并行计算能够提高算法的可扩展性,且分布式优化算法有助于保持种群多样性,避免局部最优并促进解的搜索,提高算法计算效率。因此,针对大数据特点,进一步研究如何设计和实现包括 Hadoop、Spark 等多种大数据平台下的较新的计算智能算法,以满足日益增长的大数据要求,仍将是探索大数据分析方法的重要问题。

(4)*建立先进的大数据平台*　大数据分析适用于各个领域,目前基于大数据分析的方法和技术并不成熟,针对某个具体领域,研究新的适用于解决该领域的先进大数据平台十分重要。如针对分布式优化算法,底层平台的选择会影响任务的计算时间。因此,充分利用机器学习、计算智能算法的分布、并行特点,构建新的适合求解大数据问题的先进平台非常关键。

(5)*优化分布式集群环境*　由于迭代计算广泛存在于机器学习、计算智能算法中,但 MapReduce 框架并不直接支持迭代算法,如何改变 Hadoop 分布式平台中的节点数量,如何优化 MapReduce 分布式编程中 Map、Reduce 的设计等来提升 MapReduce 对迭代计算任务的性能,仍然值得进一步深入研究。

(6)*分配深度神经网络的并行训练*　Spark 采用主从架构进行分布式计算,但不是所有的深度神经网络模型都能部署在分布式 Spark 集群上进行并行处理。且随着深度学习规模的逐步扩大,还需要考虑节点的数量、通信效率与负载均衡等问题,如何分配深度神经网络的并行训练仍然是一个具有挑战性的问题。

(7)*优化深度学习参数*　随着深度学习深度的不断增加,参数规模逐渐扩大,面对海量参数的优化是一个大规模优化问题。运用 Spark 分布式计算平台,能够加速深度学习训练,结合智能优化算法来优化深度学习参数有助于提升深度学习算法的性能。因此,如何使用计算智能算法优化深度学习模型相关参数,进一步提高深度学习算法的效率是一个重要研究方向。

(8)*改进深度学习模型*　为满足大数据处理需求,怎样降低深度学习模型训练时间成本非常重要。但是基于深度学习的大数据分析,仍然存在许多需要进一步解决的关键问题。如,除了基于 MapReduce/Spark 分布式计算平台来降低深度学习模型的时间成本之外,是否能够通过对深度学习模型自身的改进、以及融合一些其余策略来加速深度网络的训练,怎样减少整个深度网络的通信代价都值得进一步深入研究。

7　结束语

基于人工智能的大数据分析方法的研究已取得了一些成果,本文系统综述了这方面的最新研究进展。从理论研究角度,首先,针对大数据的聚类、关联分析、分类、预测 4 种不同的挖掘任务,阐述了大数据环境下机器学习分析方法的研究现状;并针对深度学习这一热点,重点分析了基于 MapReduce、Spark 的分布式深度学习以及面向大数据分析的深度学习算法的改进;然后,从群智能、进化算法两方面梳理了基于计算智能的大数据分析相关研究,由于计算智能方法不依赖于知识而直接对数据进行分析,且由于具有分布式和并行计算的优势,非常适合进行大数据分析。

同时,本文从工程实现角度,对大数据平台进行了相关归纳。为了解决大数据的深度学习模型难以放入仅有单个 GPU 的计算机中进行运算的问题,集成大数据分析和深度学习框架再通过大规模计算集群实现分布式深度学习,有助于基于深度学习的大数据分析算法执行效率的提升。为了促进大数据分析算法的快速工程应用,国内外已经展开了对大数据机器学习系统的相关研究工作,主要有南京大学 PASA 大数据实验室的 Octopus、美国卡耐基梅隆大学的 GraphLab 和 Petuum、新加坡国立大学黄铭钧团队的 SINGA、百度的 ELF 和腾讯的 Angel

等。以及包含部署在分布式计算平台上的经典机器学习算法库，主要有 MLlib、Mahout、Oryx2、H2O、Fregata、SystemML 和 Scikit-learn 等。

此外，还对当前基于人工智能技术的大数据分析方法面临的挑战及进一步的研究方向进行了讨论。总之，人工智能技术在大数据分析中具有巨大的应用潜力，尽管目前已经有了一些探索性的理论和工程研究工作，但成果尚不丰富。从总体上看，基于人工智能技术的大数据分析研究还处于起步阶段，尚有大量具有挑战性的关键问题需要深入研究。

参考文献：

[1] LABRINIDIS A, JAGADISH H V. Challenges and opportunities with big data[J]. Proceedings of the VLDB Endowment, 2012,5(12):2032-2033.

[2] LI Xuelong, GONG Haigang. A survey on big data systems [J]. Scientia Sinica Informations, 2015, 45(1): 1-44(in Chinese).[李学龙，龚海刚. 大数据系统综述[J]. 中国科学：信息科学，2015,45(1):1-44.]

[3] WANG Wanliang. Artificial intelligence: principles and applications[M]. 3rd ed. Beijing: Higher Education Press, 2016(in Chinese).[王万良. 人工智能及其应用[M]. 3版. 北京：高等教育出版社，2016.]

[4] WANG L. Machine learning in big data[J]. International Journal of Advances in Applied Sciences, 2016, 4(4): 117-123.

[5] JAPKOWICZ N, STEFANOWSKI J. A machine learning perspective on big data analysis[M]//Big Data Analysis: New Algorithms for a New Society. Berlin, Germany: Springer-Verlag, 2016: 1-31.

[6] GRIMMER J. We are all social scientists now: how big data, machine learning, and causal inference work together[J]. Political Science & Politics, 2015, 48(1): 80-83.

[7] ZHAO W, MA H, HE Q. Parallel k-means clustering based on mapreduce[C]//Proceedings of the 1st International Conference on Cloud Computing. Berlin, Germany: Springer-Verlag, 2009: 674-679.

[8] GAO H, JIANG J, SHE L, et al. A new agglomerative hierarchical clustering algorithm implementation based on the MapReduce framework[J]. International Journal of Digital Content Technology and its Applications, 2010, 4(3): 95-100.

[9] HE Y, TAN H, LUO W, et al. Mr-dbscan: an efficient parallel density-based clustering algorithm using mapreduce[C]// Proceedings of the 17th International Conference on Parallel and Distributed Systems (ICPADS). Washington, D. C., USA: IEEE, 2011: 473-480.

[10] YAN W, BRAHMAKSHATRIYA U, XUE Y, et al. p-PIC: parallel power iteration clustering for big data [J]. Journal of Parallel and Distributed Computing, 2013, 73(3): 352-359.

[11] ZHAO Y, CHEN Y, LIANG Z, et al. Big data processing with probabilistic latent semantic analysison MapReduce [C]//Proceedings of the 2014 International Conference on Cyber Enabled Distributed Computing and Knowledge Discovery (CyberC). Washington, D. C., USA: IEEE, 2014: 162-166.

[12] KIM Y, SHIM K, KIM M S, et al. DBCURE-MR: an efficient density-based clustering algorithm for large data using MapReduce[J]. Information Systems, 2014, 42: 15-35.

[13] HU C, KANG X, LUO N, et al. Parallel clustering of big data of spatio-temporal trajectory [C]//Proceedings of the 11th International Conference on Natural Computation. Washington, D. C., USA: IEEE, 2015: 769-774.

[14] BU Fanyu, CHEN Zhikui, LI Peng, et al. A high-order CFS algorithm for clustering big data[EB/OL]. [2017-08-10]. https://www.hindawi.com/journals/misy/2016/4356127.

[15] LIAO Q, YANG F, ZHAO J. An improved parallel K-means clustering algorithm with MapReduce[C]//Proceedings of the 15th IEEE International Conference on Communication Technology. Washington, D. C., USA: IEEE, 2013: 764-768.

[16] CUI X, ZHU P, YANG X, et al. Optimized big data k-means clustering using MapReduce[J]. The Journal of Supercomputing, 2014, 70(3): 1249-1259.

[17] AKTHAR N, AHAMAD M V, KHAN S. Clustering on big data using hadoop MapReduce[C]//Proceedings of the 2015 International Conference on Computational Intelligence and Communication Networks. Washington, D. C., USA: IEEE, 2015: 789-795.

[18] XIA Dawen, WANG B, LI Yantao, et al. An efficient MapReduce-based parallel clustering algorithm for distributed traffic subarea division [EB/OL]. [2017-08-10]. https://www.hindawi.com/journals/ddns/2015/793010/abs.

[19] NGUYEN C D, NGUYEN D T, PHAM V H. Parallel two-phase K-means[C]//Proceedings of the International Conference on Computational Science and Its Applications. Berlin, Germany: Springer-Verlag, 2013: 224-231.

[20] AGRAWAL R, IMIELIŃSKI T, SWAMI A. Mining association rules between sets of items in large databases[C]//Proceedings of the ACM Sigmod Record International Conference on Management of Data. New York, N. Y., USA: ACM, 1993: 207-216.

[21] HAN J, PEI J, YIN Y. Mining frequent patterns without candidate generation [C]//Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. New York, N. Y., USA: ACM, 2000: 1-12.

[22] LI N, ZENG L, HE Q, et al. Parallel implementation of apriori algorithm based on mapreduce[C]//Proceedings of the 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel & Distributed Computing (SNPD). Washington, D. C., USA: IEEE, 2012: 236-241.

[23] EZHILVATHANI A, RAJA K. Implementation of parallel apriori algorithm on hadoop cluster[J]. International Journal of Computer Science and Mobile Computing, 2013, 2(4):

513-516.

[24] ORUGANTI S, DING Q, TABRIZI N. Exploring Hadoop as a platform for distributed association rule mining[C]// Proceedings of the F C 5th International Conferenceon Future Computational Technologies and Applications. New York, N. Y. ,USA:IARIA,2013:62-67.

[25] HAO X F, TAN Y S, WANG J Y. Research and implementation of parallel apriori algorithm on Hadoop platform[J]. Computer and Modernization,2013,3(1):1-5.

[26] XIAO L, HONGWU L, FANGFANG G, et al. A cloud security situational awareness model based on parallel apriori algorithm[J]. Applied Mechanics & Materials, 2014, 556-562:6294-6297.

[27] QIU H, GU R, YUAN C, et al. Yafim:a parallel frequent itemset mining algorithm with spark[C]//Proceedings of the 2014 IEEE International Parallel & Distributed Processing Symposium Workshops (IPDPSW). Washington, D. C., USA:IEEE,2014:1664-1671.

[28] SHE X Y, ZHANG L. Apriori parallel improved algorithm based on MapReduce distributed architecture[C]//Proceedings of the 6th IEEE International Conference on Instrumentation and Measurement, Computer, Communication and Control. Washington, D. C., USA:IEEE,2016:517-521.

[29] ZHOU X, HUANG Y. An improved parallel association rules algorithm based on MapReduce framework for big data [C]//Proceedings of the 11th International Conference on Fuzzy Systemsand Knowledge Discovery(FSKD). Washington, D. C., USA:IEEE,2014:284-288.

[30] MI Yunlong, JIANG Lin, MI Chunqiao. Rough association rules algorithm with negation under MapReduce[J]. Computer Integrated Manufacturing Systems, 2014, 20(11): 2893-2903(in Chinese).[米允龙,姜　麟,米春桥. MapReduce 环境下的否定粗糙关联规则算法[J]. 计算机集成制造系统,2014, 20(11):2893-2903.]

[31] PADILLO F, LUNA J M, VENTURA S. Mining perfectly rare itemsets on big data:an approach based on apriori-inverse and MapReduce[C]//Proceedings of the International Conference on Intelligent Systems Design and Applications. Berlin, Germany:Springer-Verlag,2016:508-518.

[32] FENG D, ZHU L, ZHANG L. Research on improved apriori algorithm based on MapReduce and Hbase[C]//Proceedings of the Advanced Information Management, Communicates, Electronicand Automation Control Conference (IMCEC). Washington, D. C., USA:IEEE,2016:887-891.

[33] SINGH S, GARG R, MISHRA P. Review of apriori based algorithms on mapreduce framework[C]//Proceedings of the International Conference on Communication and Computing. Washington, D. C., USA:IEEE,2014:593-604.

[34] ZHOU L, ZHONG Z, CHANG J, et al. Balanced parallel fp-growth with mapreduce [C]//Proceedings of the 2010 IEEE Youth Conference on Information Computing and Telecommunications (YC-ICT). Washington, D. C., USA: IEEE,2010:243-246.

[35] XIAO T, YUAN C, HUANG Y. PSON:A parallelized SON algorithm with MapReduce for mining frequent sets[C]// Proceedings of the 4th International Symposium on Parallel Architectures, Algorithms and Programming. Washington, D. C., USA:IEEE,2011:252-257.

[36] WANG Jie, DAI Qinghao, ZENG Yu, et al. Parallel frequent pattern growth algorithm optimizationin cloud manufacturing environment[J]. Computer Integrated Manufacturing Systems,2012,18(9):2124-2129(in Chinese).[王　洁, 戴清灏,曾　宇,等. 云制造环境下并行频繁模式增长算法优化[J]. 计算机集成制造系统,2012,18(9):2124-2129.]

[37] WANG L, FENG L, ZHANG J, et al. An efficient algorithm of frequent itemsets mining based onmapreduce[J]. Journal of Information & Computational Science, 2014, 11 (8):2809-2816.

[38] LIU Zhiyong. Parallelizable algorithms research of association rules mining[D]. Nanjing: Southeast University, 2016 (in Chinese).[刘智勇. 关联规则挖掘的并行化算法研究[D]. 南京:东南大学,2016.]

[39] XIA D, RONG Z, ZHOU Y, et al. A novel parallel algorithm for frequent itemsets mining in massive small files datasets[J]. ICIC Express Letters, Part B:Applications,2014,5 (2):459-466.

[40] BIN Z, WENSHENG X. An improved algorithm for high-speed train's maintenance data mining based on MapReduce [C]//Proceedings of the 2015 International Conference on Cloud Computing and Big Data(CCBD). Washington, D. C., USA:IEEE,2015:59-66.

[41] BECHINI A, MARCELLONI F, SEGATORI A. A MapReduce solution for associative classification of big data[J]. Information Sciences,2016,332(C):33-55.

[42] DOU Meng, WEN Lijie, WANG Jianmin, et al. Parallel algorithm to convert big event log based on MapReduce[J]. Computer Integrated Manufacturing Systems, 2013, 19(8): 1784-1793(in Chinese).[窦　蒙,闻立杰,王建民,等. 基于MapReduce的海量事件日志并行转化算法[J]. 计算机集成制造系统,2013,19(8):1784-1793.]

[43] SUN Z Y, TSAI M C, TSAI H P. Mining uncertain sequence data on hadoop platform[C]//Proceedings of the Pacific Asia Conference on Knowledge Discovery and Data Mining. Cham, Switzerland: Springer International Publishing, 2014:204-215.

[44] AGBEHADJI I E, FONG S, MILLHAM R. Wolf search algorithm for numeric association rule mining[C]//Proceedings of the 2016 IEEE International Conference on Cloud Computing and Big Data Analysis(ICCCBDA). Washington, D. C., USA:IEEE,2016:146-151.

[45] NGUYEN D, NGUYEN L T, VO B, et al. Efficient mining of class association rules with the itemset constraint[J]. Knowledge-Based Systems,2016,103(C):73-88.

[46] LEE D, QUADRIFOGLIO L, TEULADA E S D, et al. Discovering relationships between factors of round-trip car sharing by using association rules approach[J]. Procedia Engi-

neering,2016,161:1282-1288.

[47] WENG J, ZHU J Z, YAN X, et al. Investigation of work zone crash casualty patterns using association rules[J]. Accident Analysis & Prevention,2016,92:43-52.

[48] DEL R S, LÓPEZ V, BENITEZ J M, et al. On the use of MapReduce for imbalanced big data using Random Forest[J]. Information Sciences,2014,285(C):112-137.

[49] SINGH K, GUNTUKU S C, THAKUR A, et al. Big data analytics framework for peer-to-peer botnet detection using random forests[J]. Information Sciences, 2014, 278(C): 488-497.

[50] LÓPEZ V, DEL R 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(C):5-38.

[51] HUANG G, HUANG G B, SONG S, et al. Trends in extreme learning machines: A review[J]. Neural Networks, 2015,61(C):32-48.

[52] KAMAL S, RIPON S H, DEY N, et al. A MapReduce approach to diminish imbalance parametersfor big deoxyribonucleic acid dataset[J]. Computer Methods and Programs in Biomedicine,2016,131(C):191-206.

[53] KUMAR M, RATH N K, RATH S K. Analysis of microarray leukemia data using an efficient MapReduce-based K-nearest-neighbor classifier[J]. Journal of Biomedical Informatics,2016,60(C):395-409.

[54] FERNANDEZ-DELGADO M, CERNADAS E, BARRO S, et al. Do we need hundreds of classifiers to solve real world classification problems[J]. Journal of Machine Learning Research,2014,15(1):3133-3181.

[55] HAFEZ M M, SHEHAB M E, EL FAKHARANY E, et al. Effective selection of machine learning algorithms for big data analytics using apache spark[C]//Proceedings of the International Conference on Advanced Intelligent Systems and Informatics. Berlin,Germany:Springer-Verlag,2016:692-704.

[56] RUTA D. Automated trading with machine learning on big data[C]//Proceedings of the 2014 IEEE International Congress on Big Data. Washington, D. C., USA: IEEE, 2014: 824-830.

[57] SUTHAHARAN S. Big data classification: Problems and challenges in network intrusion prediction with machine learning[J]. ACM SIGMETRICS Performance Evaluation Review,2014,41(4):70-73.

[58] RAMAKRISHNAN R, DRAL P O, RUPP M, et al. Big data meets quantum chemistry approximations: the Δ-machine learning approach[J]. Journal of Chemical Theory and Computation,2015,11(5):2087-2096.

[59] GINSBERG J, MOHEBBI M H, PATEL R S, et al. Detecting influenza epidemics using search engine query data[J]. Nature,2009,457(7232):1012-1014.

[60] BIBAULT J E, GIRAUD P, BURGUN A. Big data and machine learning in radiation oncology: state of the art and future prospects[J]. Cancer Letters,2016,382(1):110-117.

[61] ZHU X, YAO J, ZHU F, et al. Wsisa: Making survival prediction from whole slide histopathological images[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington, D. C., USA: IEEE, 2017:7234-7242.

[62] SIMMHAN Y, AMAN S, KUMBHARE A, et al. Cloud-based software platform for big data analytics in smart grids [J]. Computing in Science & Engineering, 2013, 15(4): 38-47.

[63] YANG Jungang, ZHANG Jie, QIN Wei, et al. Big data analysis platform for semiconductor manufacturing[J]. Computer Integrated Manufacturing Systems,2016,22(12):2900-2910(in Chinese).[杨俊刚,张 洁,秦 威,等.面向半导体制造的大数据分析平台[J].计算机集成制造系统,2016,22(12):2900-2910.]

[64] ZHANG Jie, GAO Liang, QIN Wei, et al. Big-Data-driven operational analysis and decision-making methodology in intelligent workshop[J]. Computer Integrated Manufacturing Systems,2016,22(5):1220-1228(in Chinese).[张 洁,高亮,秦 威,等.大数据驱动的智能车间运行分析与决策方法体系[J].计算机集成制造系统,2016,22(5):1220-1228.]

[65] LYU Youlong, ZHANG Jie. Big-data-based technical framework of smart factory[J]. Computer Integrated Manufacturing Systems,2016,22(11):2691-2697(in Chinese).[吕佑龙,张 洁.基于大数据的智慧工厂技术框架[J].计算机集成制造系统,2016,22(11):2691-2697.]

[66] YAO Xifan, ZHOU Jiajun, ZHANG Cunji, et al. Procative manufacturing—a big-data driven emerging manufacturing paradigm[J]. Computer Integrated Manufacturing Systems, 2017,23(1):172-185(in Chinese).[姚锡凡,周佳军,张存吉,等.主动制造——大数据驱动的新兴制造范式[J].计算机集成制造系统,2017,23(1):172-185.]

[67] ZHU Xuechu, QIAO Fei. Cycle time prediction method of wafer fabrication system based on industrial big data[J]. Computer Integrated Manufacturing Systems,2017,23(10): 2172-2179(in Chinese).[朱雪初,乔 非.基于工业大数据的晶圆制造系统加工周期预测方法[J].计算机集成制造系统,2017,23(10):2172-2179.]

[68] WAMBA S F, GUNASEKARAN A, AKTER S, et al. Big data analytics and firm performance: effects of dynamic capabilities[J]. Journal of Business Research, 2017, 70(C): 356-365.

[69] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation,2006, 18(7):1527-1554.

[70] ZHANG C Y, CHEN C P, CHEN D, et al. MapReduce based distributed learning algorithm for restricted boltzmann machine[J]. Neurocomputing,2016,198(C):4-11.

[71] ZHANG K, CHEN X W. Large-scale deep belief nets with mapreduce[J]. IEEE access,2014,2:395-403.

[72] WANG Dewen, SUN Zhiwei. Big data analysis and parallel load forecasting of electric power userside[J]. Proceedings of the CSEE,2015,35(3):527-537(in Chinese).[王德文,孙志

伟.电力用户侧大数据分析与并行负荷预测[J].中国电机工程学报,2015,35(3):527-537.]

[73] ZHANG H J, XIAO N F. Parallel implementation of multilayered neural networks based on MapReduce on cloud computing clusters[J]. Soft Computing,2016,20(4):1471-1483.

[74] CAO J, CUI H, SHI H, et al. Big Data:a parallel particle swarm optimization-back-propagation neural network algorithm based on MapReduce [J]. PLos One, 2016, 11(6):e0157551.

[75] MAO Guojun, HU Dianjun, XIE Songyan. Models and algorithms for classifying big data based on distribued data streams[J]. Chinese Journal of Computers,2017,40(1):161-175(in Chinese).[毛国君,胡殿军,谢松燕.基于分布式数据流的大数据分类模型和算法[J].计算机学报,2017,40(1):161-175.]

[76] HE Jieyue, MA Bei. Based on real-valued conditional restricted boltzman machine and social network for collaborative filtering[J]. Chinese Journal of Computers,2016,39(1):183-195(in Chinese).[何洁月,马　贝.利用社交关系的实值条件受限玻尔兹曼机协同过滤推荐算法[J].计算机学报,2016,39(1):183-195.]

[77] LI H, SU P, CHI Z, et al. Image retrieval and classification on deep convolutional SparkNet[C]//Proceedings of the 2016 IEEE International Conference on Signal Processing,Communications and Computing (ICSPCC). Washington, D. C., USA:IEEE,2016:1-6.

[78] YANG Jiaju. MapReduce-based and deep learning for load analysis and forecasting[D]. Nanjing:Southeast University, 2016(in Chinese).[杨佳驹.基于 MapReduce 和深度学习的负荷分析与预测[D].南京:东南大学,2016.]

[79] ALSHEIKH M A, NIYATO D, LIN S, et al. Mobile big data analytics using deep learning and apache spark[J]. IEEE network,2016,30(3):22-29.

[80] OUYANG X, ZHANG C, ZHOU P, et al. DeepSpace:an online deep learning framework for mobile big data to understand human mobility patterns[EB/OL].(2017-07-09)[2018-03-05]. https://arxiv.org/pdf/1610.07009.pdf.

[81] YAN Y, CHEN M, SADIQ S, et al. Efficient imbalanced multimedia concept retrieval by deep learning on spark clusters[J]. International Journal of Multimedia Data Engineering and Management,2017,8(1):1-20.

[82] YANG Y, ZHAN D C, FAN Y, et al. Deep learning for fixed model reuse[C]//Proceedings of the AAAI. Palo Alto, Cal., USA:AAAI,2017:2831-2837.

[83] JIE Z, WEI Y, JIN X, et al. Deep self-taught learning for weakly supervised object localization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington, D. C., USA: IEEE, 2017: 4294-4302.

[84] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [EB/OL]. (2017-07-09)[2018-03-05]. https://arxiv.org/pdf/1502.03167.pdf.

[85] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Washington, D. C., USA:IEEE,2016:770-778.

[86] KRANJC J, ORAČ R, PODPEČAN V, et al. ClowdFlows: online workflows for distributed big data mining[J]. Future Generation Computer Systems,2017,68(C):38-58.

[87] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human level control through deep reinforcement learning[J]. Nature,2015,518(7540):529-533.

[88] XUE H, LIU Y, CAI D, et al. Tracking people in RGBD videos using deep learning and motion clues[J]. Neurocomputing,2016,204(C):70-76.

[89] ZHAO Peng, ZHOU Zhihua. Distribution-Free One-Pass Learning[EB/OL]. (2017-07-09)[2018-03-05]. http://www.doc88.com/p-3495610983591.html.

[90] KE W, CHEN J, JIAO J, et al. SRN:Side-output residual network for object symmetry detection in the wild[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, D. C., USA:IEEE,2017:302-310.

[91] ZHOU Y, YE Q, QIU Q, et al. Oriented response networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, D. C., USA: IEEE,2017:4961-4970.

[92] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//Proceedings of the 34th International Conference on Machine Learning. New York,N. K.,USA:ACM,2017:1126-1135.

[93] CHENG S, SHI Y, QIN Q, et al. Swarm intelligence in big data analytics[C]//Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning. Berlin, Germany:Springer-Verlag,2013:417-426.

[94] MCNABB A W, MONSON C K, SEPPI K D. Parallel-pso using mapreduce[C]//Proceedings of the 2007 IEEE Congress on Evolutionary Computation. Washington, D. C., USA: IEEE,2007:7-14.

[95] LIANG J, WANG F, DANG C, et al. An efficient rough feature selection algorithm with a multi granulation view[J]. International Journal of Approximate Reasoning, 2012, 53(6):912-926.

[96] LI X, YAO X. Cooperatively coevolving particle swarms for large scale optimization[J]. IEEE Transactions on Evolutionary Computation,2012,16(2):210-224.

[97] XU Zongben, ZHANG Wei, LIU Lei, et al. The scientific-principle and prospect for data scienceand big data the 462nd expert speech act of academic forum on Xiangshan science conference[J]. Science & Technology for Development, 2014,10(1):66-75(in Chinese).[徐宗本,张　维,刘　雷,等."数据科学与大数据的科学原理及发展前景"—香山科学会议第 462 次学术讨论会专家发言摘登[J].科技促进发展,2014,10(1):66-75.]

[98] GOVINDARAJAN K, SOMASUNDARAM T S, KUMAR

V S. Continuous clustering in big data learning analytics [C]//Proceedings of the 5th International Conference on Technology for Education. Washington, D. C., USA:IEEE, 2013:61-64.

[99] GUPTA S L, GOEL S, BAGHEL A S. An approach to handle big data analytics using potential of swarm intelligence [C]//Proceedings of the 3rd International Conference on Computing for Sustainable Global Development. Washington, D. C., USA:IEEE,2016:3640-3644.

[100] WANG L, WANG Y, CHANG Q. Feature selection methods for big data bioinformatics: A survey from the search perspective[J]. Methods,2016,111(C):21-31.

[101] CHENG S, ZHANG Q, QIN Q. Big data analytics with swarm intelligence[J]. Industrial Management & Data Systems,2016,116(4):646-666.

[102] FONG S, WONG R, VASILAKOS A V. Accelerated PSO swarm search feature selection for data stream mining big data[J]. IEEE Transactions on Services Computing,2016,9(1):33-45.

[103] CAO B, LI W, ZHAO J, et al. Spark-based parallel cooperative co-evolution particle swarm optimization algorithm [C]//Proceedings of the 2016 IEEE International Conference on Web Services. Washington, D. C., USA: IEEE, 2016:570-577.

[104] WANG Y, LI Y, CHEN Z, et al. Cooperative particle swarm optimization using MapReduce[J]. Soft Computing, 2017,21(22):6593-6603.

[105] LI Y, CHEN Z, WANG Y, et al. Quantum-behaved particle swarm optimization using mapreduce[C]//Proceedings of the Bio-Inspired Computing-Theories and Applications. Berlin, Germany:Springer-Verlag,2016:173-178.

[106] DING W, LIN C T, CHEN S, et al. Multiagent-consensus-mapreduce-based attribute reduction using co-evolutionary quantum PSO for big data applications[J]. Neurocomputing,2018,272(C):136-153.

[107] CHENG X, XIAO N. Parallel implementation of dynamic positive and negative feedback aco with iterative mapreduce model[J]. Journal of Information & Computational Science, 2013,10(8):2359-2370.

[108] WU Hao, NI Zhiwei, WANG Huiying. MapReduce-based ant colony optimization[J]. Computer Integrated Manufacturing Systems,2012,18(7):1503-1509(in Chinese).[吴昊,倪志伟,王会颖. 基于 MapReduce 的蚁群算法[J]. 计算机集成制造系统,2012,18(7):1503-1509.]

[109] MA Wenlong, WANG Zheng, ZHAO Yanwei. Optimizing services composition in cloud manufacturing based on improved ant colony algorithm[J]. Computer Integrated Manufacturing Sysems,2016,22(1):113-121(in Chinese).[马文龙,王 铮,赵燕伟. 基于改进蚁群算法的制造云服务组合优化[J]. 计算机集成制造系统,2016,22(1):113-121.]

[110] LIN C Y, PAI Y M, TSAI K H, et al. Paralllizing modified cuckoo search on mapreduce architecture[J]. Journal of Electronic Science and Technology,2013,11(2):115-123.

[111] XU X, JI Z, YUAN F, et al. A novel parallel approach of cuckoo search using MapReduce[C]//Proceedings of the 2014 International Conference on Computer, Communications and InformationTechnology (CCIT 2014). Amsterdam, the Netherlands:Atlantis Press,2014:114-117.

[112] AL-MADI N, ALJARAH I, LUDWIG S A. Parallel glowworm swarm optimization clustering algorithm based on MapReducee[C]//Proceedings of the 2014 IEEE Symposium on Swarm Intelligence. Washington, D. C., USA:IEEE,2014:1-8.

[113] ZHENG Hongsheng, YU Dongjin, ZHANG Lei. Multi-Qos cloud workflow scheduling based on firely algorithm and dynamic priorities[J]. Computer Integrated Manufacturing Systems,2017,23(5):963-971(in Chinese).[郑宏升,俞东进,张 蕾. 基于萤火虫算法和动态优先级的多 QoS 云工作流调度[J]. 计算机集成制造系统,2017,23(5):963-971.]

[114] LIN K C, ZHANG K Y, HUANG Y H, et al. Feature selection based on an improved cat swarm optimization algorithm for big data classification[J]. The Journal of Supercomputing,2016,72(8):3210-3221.

[115] JIN C, VECCHIOLA C, BUYYA R. MRPGA:an extension of MapReduce for parallelizing genetic algorithms[C]//Proceedings of the 4th International Conference on eScience 2008. Washington, D. C., USA:IEEE,2008:214-221.

[116] YANG Z, TANG K, YAO X. Large scale evolutionary optimization using cooperative coevolution[J]. Information Sciences,2008,178(15):2985-2999.

[117] OMIDVAR M N, LI X, MEI Y, et al. Cooperative coevolution with differential grouping for large scale optimization [J]. IEEE Transactions on Evolutionary Computation, 2014,18(3):378-393.

[118] QIAN Chao, ZHOU Zhihua. Decomposition-based pareto-optimization for subset selection[J]. Scientia Sinica Informations,2016,46(9):1276-1287(in Chinese).[钱 超,周志华. 基于分解策略的多目标演化子集选择算法[J]. 中国科学:信息科学,2016,46(9):1276-1287.]

[119] GHEYAS I A, SMITH L S. Feature subset selection in large dimensionality domains[J]. Pattern Recognition, 2010,43(1):5-13.

[120] BACARDIT J, LLOR X. Large-scale data mining using genetics-based machine learning[J]. Wiley Inter Disciplinary Reviews: Data Mining and Knowledge Discovery, 2013, 3(1):37-61.

[121] YUAN F, LIAN F, XU X, et al. Decision tree algorithm optimization research based on MapReduce[C]//Proceedings of the 6th IEEE International Conference on Software Engineering and Service Science. Washington, D. C., USA:IEEE,2015:1010-1013.

[122] ZHANG Ying, ZHAI Li, WANG Jing. Cloud computing federation data resource server comppsotion in big data background[J]. Computer Integrated Manufacturing Systems,2016,22(12):2920-2929(in Chinese).[张 影,翟丽丽,王 京. 大数据背景下的云联盟数据资源服务组合模型

[J]. 计算机集成制造系统,2016,22(12):2920-2929.]

[123] ZHU Linan, WANG Wanliang, SHEN Guojiang. Resource optimization combinaton method basedon improved differential evolution algorithm forcloud manufaturing[J]. Computer Integrated Manufacturing Systems, 2017, 23(1): 203-214 (in Chinese). [朱李楠,王万良,沈国江. 基于改进差分进化算法的云制造资源优化组合方法[J]. 计算机集成制造系统, 2017,23(1):203-214.]

[124] MORITZ P, NISHIHARA R, STOICA I, et al. Sparknet: training deep networks in spark [EB/OL]. (2017-07-09) [2018-03-05]. https://arxiv.org/pdf/1511.06051.pdf.

[125] Computer Network Information Center, Chinese Academy of Sciences. Depth learning method and system for big data: China, CN106570565 A[P/OL]. (2017-04-19)[2017-08-28]. https://www.google.com/patents/CN106570565A?cl=en&hl=zh-CN(in Chinese). [中国科学院计算机网络信息中心. 一种面向大数据的深度学习方法及系统:中国, CN106570565 A[P/OL]. (2017-04-19)[2017-08-28]. https://www.google.com/patents/CN106570565A? cl = en&hl=zh-CN.]

[126] KIM H, PARK J, JANG J, et al. DeepSpark: a spark-based distributed deep learning framework for commodity clusters [EB/OL]. (2017-07-09)[2018-03-05]. https://arxiv.org/pdf/1602.08191.pdf.

[127] TEAM D. Deeplearning4j: Open-source distributed deep learning for the JVM. [EB/OL]. [2017-08-28]. https://deeplearning4j.org/.

[128] LIU S, TANG J, WANG C, et al. Implementing a cloud platform for autonomous driving [EB/OL]. (2017-07-09) [2018-03-05]. https://arxiv.org/ftp/arxiv/papers/1704/1704.02696.pdf.

[129] XING E P, HO Q, XIE P, et al. Strategies and principles of distributed machine learning on big data[J]. Engineering, 2016,2(2):179-195.

[130] HUANG Yihua. Research progress on big data machine learning system[J]. Big Data Research, 2015, 1(1): 28-47(in Chinese). [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.]

[131] LOW Y, BICKSON D, GONZALEZ J, et al. Distributed GraphLab: a framework for machine learning and data mining in the cloud[J]. Proceedings of the VLDB Endowment, 2012,5(8):716-727.

[132] XING E P, HO Q, DAI W, et al. Petuum: a new platform for distributed machine learning on big data [J]. IEEE Transactions on Big Data, 2015, 1(2): 49-67.

[133] OOI B C, TAN K L, WANG S, et al. SINGA: a distributed deep learning platform[C]//Proceedings of the 23rd ACM international conference on Multimedia. New York, N. Y., USA: ACM, 2015: 685-688.

[134] LI M, ANDERSEN D G, PARK J W, et al. Scaling distributed machine learning with the parameter server[C]//Proceedings of the OSDI. Berkeley, Cal., USA: USENIX Association, 2014: 583-598.

[135] JIANG J, YU L, JIANG J, et al. Angel: a new large-scale machine learning system [J]. National Science Review, 2017: 5(2): 216-236.

[136] POP D, IUHASZ G, PETCU D. Distributed platforms and cloud services: enabling machine learning for big data[M]// Data Science and Big Data Computing. Cham, Switzerland: Springer International Publishing, 2016: 139-159.

[137] MENG X, BRADLEY J, YAVUZ B, et al. MLlib: machine learning in apache spark[J]. The Journal of Machine Learning Research, 2016, 17(1): 1235-1241.

[138] OWEN S, ANIL R, DUNNING T, et al. Mahout in action [M]. New York, N. Y., USA: Manning Publications, 2011.

[139] DINO K. H2O persistence framework for column oriented distributed(NoSQL) databases[C]//Proceedings of the 3rd International Symposiu on Sustainable Development. Southampton, UK: Eprints, 2012: 22-28.

[140] ZHANG X, YAO F, TIAN Y. Greedy step averaging: a parameter-free stochastic optimization method [EB/OL]. (2016-11-14)[2018-03-15]. http://pdfs.semanticscholar.org/256d/e07d93d8e4021b444e0c6905997fc64e9d70.pdf.

[141] GHOTING A, KRISHNAMURTHY R, PEDNAULT E, et al. SystemML: declarative machine learning on MapReduce[C]//Proceedings of the 27th International Conference on Data Engineering (ICDE). Washington, D. C., USA: IEEE, 2011: 231-242.

[142] PEDREGOSA F, VAROQUAUX G, GRAMFORT A, et al. Scikit-learn: machine learning in python[J]. Journal of Machine Learning Research, 2011, 12: 2825-2830.

[143] DEAN J, CORRADO G, MONGA R, et al. Large scale distributed deep networks[C]//Advances in Neural Information Processing Systems. Cambridge, Mass., USA: MIT Press, 2012: 1223-1231.

[144] GONG Y J, CHEN W N, ZHAN Z H, et al. Distributed evolutionary algorithms and their models: a survey of the state-of-the-art[J]. Applied Soft Computing, 2015, 34(C): 286-300.

作者简介:

王万良(1957—),男,江苏高邮人,教授,博士生导师,国家教学名师,国家万人计划领军人才,全国大数据教育联盟副理事长,浙江省人工智能学会副理事长,研究方向:人工智能、大数据分析、优化调度、计算机智能自动化等,E-mail: wwl@zjut.edu.cn;

张兆娟(1990—),女,江西九江人,博士研究生,研究方向:大数据、优化调度、深度学习;

高　楠(1983—),女,安徽合肥人,讲师,研究方向:大数据、深度学习、机器学习、生物信息;

赵燕伟(1959—),女,河南郑州人,教授,博士生导师,研究方向:先进制造、数字化装备设计、物流系统建模与优化。