工业大数据分析综述：模型与算法

王宏志，梁志宇，李建中，高宏
哈尔滨工业大学计算机科学与技术学院，黑龙江 哈尔滨 150001

摘要
随着条形码、二维码、RFID、工业传感器、自动控制系统、工业互联网、ERP、CAD/CAM/CAE等信息技术在工业领域的广泛应用，大量与工业生产活动相关的数据被实时采集并存储到企业的信息系统中。对这些数据进行分析，有助于改进生产工艺、提高生产效率、降低生产成本，为实现智能制造奠定基础。因此，工业大数据分析引起了工业界和学术界的广泛关注。模型和算法是大数据分析理论和技术中的两个核心问题。介绍了工业大数据分析的基本概念，综述了几种流行的工业大数据分析模型在工业大数据分析领域的应用情况以及相应求解算法方面的研究成果，并探索了大数据分析模型和算法的未来研究方向。

关键词
工业大数据；大数据分析；模型；算法；智能制造

中图分类号：TP31 文献标识码：A doi: 10.11959/j.issn.2096-0271.2018051

Survey on industrial big data analysis: models and algorithms

WANG Hongzhi, LIANG Zhiyu, LI Jianzhong, GAO Hong
School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China

Abstract
With the wide application of bar code, two-dimensional code, RFID, industrial sensor, automatic control system, industrial Internet, ERP and CAD/CAM/CAE techniques in industry, a large number of data related to industrial production are collected and stored in information system in real time. Analyzing those data can help to improve the production techniques, optimize the production process, reduce the production costs, laying the foundation for intelligent manufacturing. Therefore, the industrial big data analysis has drawn much attention of both industry and academia in recent years. Models and algorithms are two core issues of big data analysis theory and techniques. The concept of industrial big data analysis was introduced, and the applications of several popular models and the research results of the corresponding algorithms were reviewed, and future research directions in this area were explored.

Key words
industrial big data, big data analysis, model, algorithm, intelligent manufacturing

1 引言

当今时代，信息化和工业化的融合已经成为发展趋势，《中国制造2025》指出：“新一代信息技术与制造业深度融合，正在引发影响深远的产业变革，形成新的生产方式、产业形态、商业模式和经济增长点”。工业大数据在两化融合过程中起着至关重要的作用，国务院颁发的《促进大数据发展行动纲要》把发展工业大数据列为主要任务之一：“推动大数据在工业研发设计、生产制造、经营管理、市场营销、售后服务等产品全生命周期、产业链全流程各环节的应用，分析感知用户需求，提升产品附加价值，打造智能工厂。建立面向不同行业、不同环节的工业大数据资源聚合和分析应用平台”。

工业大数据是指在工业领域中产生的大数据。随着信息化与工业化的深度融合，信息技术渗透到了工业企业产业链的各个环节，条形码、二维码、射频识别（radio frequency identification，RFID）、工业传感器、工业自动控制系统、工业互联网、企业资源计划（enterprise resource planning，ERP）、计算机辅助设计（computer aided design，CAD）、计算机辅助制造（computer aided manufacturing，CAM）、计算机辅助工程（computer aided engineering，CAE）等技术在工业企业中得到广泛应用。互联网、移动互联网、物联网等新一代信息技术在工业领域的应用使得工业企业也进入新的发展阶段，其拥有的数据也日益丰富。尤其是制造企业中生产线处于高速运转的状态，工业设备上产生了大量数据，同时，企业中人和计算机也产生了大规模数据。

例如，通用电气公司（GE）能源监测和诊断中心每天从客户那里收集10 GB的数据[1]，长虹集团有限公司等离子显示板生产流程数据涉及超过10 000个参数，每天产生3 000万条记录，数据量大约为10 GB[2]。杭州西奥电梯有限公司的数字化车间监控超过500个参数，每天产生约50万条记录；浙江雅莹服装有限公司数字化生产线由15个子系统组成，每天产生约80万条记录，数据量大约为1 GB。

模型和算法是大数据分析中的两个核心问题。大数据分析模型的研究可以分为3个层次，即描述分析、预测分析和规范分析。描述分析探索历史数据，并描述发生了什么，这一层次包括发现数据规律的聚类[3]、相关规则挖掘[4]、模式发现[5]和描述数据规律的可视化分析[6]；预测分析用于预测未来的概率和趋势，例如基于逻辑回归的预测[7]、基于分类器的预测[8]等；规范分析根据期望的结果、特定场景、资源以及对过去和当前事件的了解对未来的决策给出建议，例如基于模拟的复杂系统分析[9]和基于给定约束的优化解生成[10]。大数据分析算法的研究则针对特定的分析模型设计高效的算法，并研究如何提高算法的可扩展性、实时性等。工业大数据除了具有大数据的5V（volume、velocity、variety、value、veracity）特征，还具有诸多与工业生产特性密切相关的特征，如数据来源多样、数据质量低、数据蕴含信息复杂、耦合不确定、数据实时性高等，这使得工业大数据分析的模型更加复杂多样，对算法的实时性要求更高。

目前，工业大数据分析的理论和技术研究仍处于起步阶段，分析模型以简单的单一模型为主，分析算法主要是通用的数据挖掘算法。本文对工业大数据分析模型和算法的研究现状进行了综述，总结了大

数据分析模型在工业场景中的应用，并对工业大数据分析模型的常用求解算法进行概述和评价。

2 大数据分析模型在工业中的应用

大数据分析模型按照输入输出的形式，主要有关联规则分析、分类分析、回归分析和聚类分析，这些分析模型以及它们的组合被应用于多种实际工业场景的分析任务中，如工艺参数优化、产量预测、故障检测和诊断、客户需求分析和服务类型识别等。

2.1 关联规则分析模型

关联规则挖掘是针对购物篮分析的问题提出的，其最初的目的是挖掘交易数据库中不同商品间存在的关联关系，以此得到顾客购买模式的一般性规则，并用这些规则指导商家进行合理的货架设计[11]。在实际工业过程中，许多场景可以用类似的关联关系建模，用频繁项集挖掘的方法来获取隐含规则，用数据和知识驱动的方式来代替传统的大量依靠经验的决策方式。

在工艺参数优化方面，参考文献[12]针对钢铁制造工艺，用关联规则分析制造过程中的化学成分（如碳、锰、磷、硫等）的含量、出钢温度和轧制速度对产品机械性能的影响。

在能耗优化方面，参考文献[13]用关联规则分析从大量历史能耗数据中获取生产参数对能耗影响的规则，以此来指导企业改进生产参数，优化能耗。

在工艺标准优化方面，参考文献[14]用关联规则分析服装款式、材料、年龄对实际工时与标准工时差异的影响程度，作为标准工时优化的依据；参考文献[15]用关联规则模型评估导致标准工时不合理的影响因素；参考文献[16]用关联规则挖掘模具企业计算机辅助工艺过程设计（computer aided process planning，CAPP）改模知识库，以此指导工艺设计，提高改模工艺继承性与重用性，促进改模工艺标准化。

在产品质量优化方面，参考文献[17]关注装配顺序对产品质量的影响。首先用关联规则的方法从历史数据中抽取规则，再经人工整理，获得简洁有效的装配知识和规则，以此指导产品按合理顺序装配。

在企业的供应链管理方面，参考文献[18]用关联规则分析机械制造企业外购件间的隐含关系，指导企业制定采购计划；参考文献[19]用关联规则模型分析汽车企业产品价格、供应商评价、退货原因间的关联关系，以此建立采购决策支持系统；参考文献[20]用关联规则模型评价和选择供应商。

在销售和售后方面，参考文献[21]从卡车制造商的角度出发，用关联规则分析经销商订购车型间的关系，以进行品牌推广；参考文献[22]将关联规则运用到笔记本电脑接单制造（build to order，BTO）生产计划中，挖掘用户配置选择的关联关系，从用户和市场需求的角度出发，为生产计划制定提供合理依据；参考文献[23]用关联规则对钟表企业物料清单中的历史数据建模，分析不同零部件组合的产品的客户认可度和盈利能力，以此为基础制定营销策略；参考文献[24]在客户关系管理（customer relationship management，CRM）中运用关联规则来预测客户行为；参考文献[25]用关联规则分析通信设备制造业不同地区的销售特征，预测流失客户。

除了直接用关联规则模型进行工业大数据分析之外，也有研究者用关联规则对工业数据进行预处理，再结合其他方法对处理过的数据进行分析。参考文献[26]针对半导体制造工艺过程中机器故障由多因素引起、因素间非线性的特点，提出了RMI（root-cause machine identifier）综合处理程序。其关键步骤之一就是用关联规则模型对数据进行处理，得到候选集，再在候选集上进行计算来定位和发现机器缺陷；参考文献[27]用关联规则对毛毯制造数据进行降维，找到关键属性集，再利用关键属性集分析毛毯制造过程中出现故障的原因和相应的规则，帮助识别和定位故障。

2.2 分类模型和回归模型

分类模型和回归模型在基于大数据的分析、判断和预测领域具有广泛应用。其中，分类模型主要针对离散属性值的判断和预测，如故障检测和诊断、客户细分；而回归模型主要针对连续属性值，如产品的产量、销量。决策树模型和神经网络模型是基于工业大数据的分类和回归分析中的主要模型。

决策树是一种树形结构，用来表征对象属性与对象值之间的映射关系。决策树模型简单直观、可解释性强，具有良好的分析和预测能力，适用于工业大数据分析的诸多场景。

在工艺参数优化方面，参考文献[28]用分类决策树对玻璃镀膜工艺建模，根据给定的工艺参数预测产品质量，从而帮助企业快速确定符合预期质量产品的工艺参数。在成本优化方面，参考文献[29]用决策树模型筛选对工艺成本影响最大的工序，通过控制和改进筛选出的工序优化整个工艺路线的成本。在工艺标准优化方面，参考文献[14]用决策树模型对服装标准工时系统的数据进行分类，再在每一类数据上进行挖掘。在产品质量监控方面，针对传统冷轧产品缺陷检测方法和检测标准不一、漏洞和误判多的问题，参考文献[30]使用生产数据训练二分类和多分类决策树，挖掘出定位产品缺陷的规则，以此帮助企业快速准确地检测出生产产品的缺陷。类似地，在故障检测方面，参考文献[27]用毛毯制造数据训练决策树模型，得到导致生产线故障的规则，并用这些规则指导相关人员识别和定位生产线故障，大大提高了毛毯生产的效率；参考文献[31]用提升机各系统的监测数据训练决策树，表征故障隐患。在生产调度方面，决策树模型受到了广泛关注。参考文献[32]针对离散制造业静态Job Shop调度问题，用决策树模型提取调度知识，优化调度方案；参考文献[33]将工单数据和机器加工数据作为历史数据训练决策树模型，并用训练好的模型实时为工件安排机器进行加工，充分利用生产资源，实现智能化生产；参考文献[34]用决策树模型分析机器细小位置移动次数和机器效率对停止次数的影响规律，作为机器调整和检测的依据，有助于车间设备、人员的调配。在采购方面，参考文献[35]用决策树模型挖掘汽车产品销售数据，指导管理者制定合理的采购计划。在客户关系管理方面，决策树模型被广泛用于客户的评价、细分和预测。参考文献[21]用决策树模型分析卡车经销商的重要程度；参考文献[36]用决策树模型将已有客户群按其忠诚度分类；参考文献[37]则用决策树模型来预测客户的忠诚度，以帮助企业发掘潜在客户；参考文献[24]用决策树模型挖掘客户潜在服务需求，帮助企业从被动服务模式向主动服务模式进化；参考文献[38]用决策树模型分析客户评价，以此衡量产品的满意度指

标；参考文献[39]用决策树模型预测潜在流失客户，指导企业维系客户关系。此外，在制造业企业的管理流程优化方面，参考文献[40]对机械制造企业的信息管理系统建立处理部门归属决策树模型，以规范管理流程，提高企业各部门间的协调运作效率。

神经网络模型能够表征复杂的非线性函数，在分类和回归任务上具有良好的表现。实际工业场景（如电网、生产线和大型制造设备等）的工艺过程、产品质量、能耗、故障等均受众多因素影响，影响过程非线性，且影响因素间往往存在着耦合关系。利用这些过程产生的大数据来训练神经网络，可以有效地表征这些复杂过程，实现工艺流程优化、质量管理监控、能耗优化、故障检测预警等。

在工艺流程优化方面，参考文献[28]对玻璃镀膜工艺建立神经网络模型，根据给定参数预测产品质量，再以产品质量为指标优化工艺参数；参考文献[41]用动态神经网络模型分析发动机缸盖机加工参数、模具状态和成品缸盖燃烧室容积间的关系，指导加工参数的选取；参考文献[42]用前馈神经网络模型对退火数据建模，模型可根据当前数据预测退火工序的运行情况，为相关人员优化退火工艺参数提供合理依据；类似地，参考文献[43]用前馈神经网络来量化表征树脂中间体和印刷电路板制造流程中各因素间的非线性因果关系，帮助相关人员更好地认识工艺流程运行情况。此外，参考文献[15]基于神经网络预测生产工时，分析出影响因素及其影响程度。在能耗的管理方面，参考文献[44]对烟草企业能源管理系统运用附加动量的前馈神经网络模型，监控和预测企业的能耗，保障企业的用能安全和效率；参考文献[45]利用造纸业电耗、气耗等4类能耗的重要影响因素建立神经网络模型，预测造纸过程中各工序的能耗，为企业的能耗管理提供决策依据。在故障检测方面，参考文献[46]用模块化的神经网络模型结合故障检测理论，实现快速定位故障部件；参考文献[47]用神经网络对数据仓库的数据进行建模，分析汽车售后的潜在故障；参考文献[48]基于模糊神经网络实现起重车远程实时故障诊断；参考文献[49]用神经网络模型解决设备实时故障数据库中隐形故障这类不确定性较多的非线性问题。在采购和库存管理方面，参考文献[50]用神经网络模型预测大型制造企业库存安全水平；参考文献[51]用神经网络模型预测服装企业库存量；参考文献[52]用神经网络模型预测采购价格。这些预测结果为合理制定采购计划、节约成本、优化库存提供了依据。此外，参考文献[53]用神经网络模型对制造企业的库存管理系统进行优化。神经网络模型还被用作产量、销量预测及客户需求发现。在产量预测方面，参考文献[54]用神经网络模型预测攀枝花市的钢产量；在销量预测方面，参考文献[55]以电机企业的月份因子及月销售量作为输入，建立神经网络模型，较好地预测出了5种电机产品的销售量；在客户需求发现方面，参考文献[21]从卡车制造商的角度出发，用神经网络模型综合分析经销商的价值。还有研究将神经网络模型运用到工业产品的辅助设计中，如参考文献[56]用神经网络模型分类并辅助检索摩托车总体设计方案实例库中的设计实例，有助于提高摩托车总体方案设计的效率。

此外，统计分析中的回归模型在工业大数据分析和预测上依然发挥着重要作用。参考文献[57]针对传统的基于伸缩探针检测半导体质量时，伸缩探针老化影响检测结果的问题，用线性回归结合分类模型对伸缩探针的状态建模，掌握探针的老

化程度，从而减少由于探针老化引起的产品质量检测错误，提高检测准确率；参考文献[58]基于动态改进多元线性回归模型预测航空制造企业的次年产量；参考文献[59]用回归模型挖掘制造业集群规模和创新能力与服务需求之间的关系；参考文献[60]用时间序列分析模型预测流行饰品在不同区域的未来直销趋势。

2.3 聚类分析模型

聚类分析模型能够将具有相似模式的对象归纳为一簇，是一种典型的无监督学习模型。聚类分析模型擅于从看似关系复杂、未知的对象中提取出内在联系。因此，在工业大数据分析中，聚类分析模型被用于分析复杂参数间的关系、细分客户群等。

在生产优化方面，参考文献[61]用聚类分析模型分析炼钢厂脱硫数据，得到不同铁水条件下的脱硫剂消耗，以此为依据确定不同铁水条件下脱硫效果最好的脱硫剂；参考文献[13]用聚类分析方法从大量历史炼钢数据中挖掘隐含的模式，指导炼钢企业改进生产参数，降低能耗；参考文献[62]用多元统计技术和聚类分析方法对车间历史数据进行分析，得到了工艺参数与能耗间的影响规律；参考文献[63]用聚类模型挖掘酿酒数据中隐含的模式，分析结果可以帮助研究人员在酿酒过程中提前检测出产品质量是否合格，减少由于成品质量问题带来的损失。

在异常检测方面，参考文献[64]提出将聚类分析模型应用到钢铁等冶金行业，分析并找出生产数据中的异常点，为进一步发现生产过程中的异常状况或根据正确数据进行更深层次的分析奠定基础；参考文献[65]将聚类和关联规则模型结合，基于大量数据预测挖掘机的故障；参考文献[47]用聚类模型分析汽车售后的潜在故障。

在供应链管理方面，参考文献[50]用灰色聚类法对库存中的产品和原材料按重要性进行划分，指导企业控制和优化库存；参考文献[21]从卡车制造商的角度出发，用聚类模型对经销商进行细分；参考文献[36]对烟草企业客户按价值度聚类，以此划分营销等级，实现精准营销；参考文献[66]从客户的盈利能力角度出发，用聚类模型分析客户盈利数据，并预测其未来盈利能力，决策者参考预测结果，将有限的资源投入能为公司创造更大利润的客户上。

在售后服务方面，参考文献[24]提出用聚类模型挖掘客户对服务的潜在需求，以客户需求为导向制定合理的服务策略；参考文献[39]用聚类模型结合决策树模型预测潜在流失客户，帮助企业及时制定挽留策略。

3 大数据分析算法研究成果

工业大数据分析算法是对特定分析模型的求解方法。如第2节所述，目前工业大数据分析中的模型以通用模型为主，因此，求解这些模型的算法主要是经典的大数据分析算法及其改进算法。本节对关联规则分析、决策树分析、神经网络和聚类分析的有关算法方面的主要研究成果进行整理和评价。

3.1 关联规则分析算法

关联规则分析中最经典的算法是Apriori算法和FP-Growth算法。

Apriori算法[67]的理论依据是频繁项集的两个重要性质，即一个频繁项集的任

意子集是频繁的，且一个非频繁项集的任意超集是非频繁的。其算法思想是首先扫描1次数据集，得到1-频繁项集，然后通过迭代逐层由(k_1) -频繁项集得到k-候选集，并利用频繁项集的性质从k-候选集中筛选k-频繁项集，直到没有新的频繁项集产生为止。得益于频繁项集的性质，Apriori算法相较于蛮力法，大大提高了计算效率，且算法的思想简单，因而在关联规则分析领域得到了广泛应用。然而，经典的Apriori算法存在两个主要问题：其一是当数据量很大时，算法会产生大量的候选集；其二是算法需要多次扫描数据集，具有很大的I/O开销。

针对Apriori算法存在的问题，许多研究者提出了优化和改进的方法。Mannila H等人[68]提出剪枝技术，利用“一个项集是频繁项集，当且仅当它的所有子集都是频繁项集”这一充分必要条件，排除那些存在不频繁子集的候选集，大大减少了候选集的规模；Park J S等人[69]提出了用散列的方法计算支持度计数的DHP算法，以空间开销换取时间开销；Savasere A等人[70]提出了基于划分的算法，将数据库划分为几个不相交的子库，分别将这些子库读入内存并计算频繁项集，再将计算结果合并，这样，每个子数据库只需要被扫描一次，减少了I/O开销；Mannila H等人[68]还提出了采样的思想；Toivonen H[71]设计了基于采样思想的算法，在数据库中抽取样本，并在样本上挖掘关联规则，再在剩余数据中验证这些规则，基于采样的算法大大提高了运行效率，但产生的结果不精确，存在“数据扭曲”的问题。此外，许多算法在Apriori算法的基础上实现了并行化，如Agrawal R等人[72]提出的CD算法、DD算法、CaD算法及Park J S等人[69]提出的PDM算法。

关联规则分析中另一个经典的算法是Han J等人[73]提出的FP-Growth算法。算法基于FP-tree的数据结构来压缩和表示数据，因而不需要产生候选集。算法首先扫描两次数据集，构造FP-tree，再利用分治的思想，在构造好的FP-tree上进行挖掘，不需要多次扫描数据集。当数据集中的事务在FP-tree上构成的重叠路径较多、FP-tree的规模足够小时，FP-Growth算法的运行效率比Apriori算法提高几个数量级。

3.2 决策树算法

决策树算法的核心是在决策树的每个节点上选取合适的测试属性，并按照测试属性将数据集进行划分，以此构造出完整的决策树。

最早的决策树算法是ID3算法[74]。ID3算法将信息熵理论引入决策树学习中，以信息增益为标准选取树节点的测试属性，递归地构造决策树。ID3算法思想简单，且具有较强的学习能力。不过，由于ID3算法偏向于处理具有较多值的属性，因而存在过拟合的问题；算法对噪声数据敏感，且算法只能处理离散值，无法对连续属性值进行处理。

在ID3算法的基础上，Quinlan J R[75]提出了C4.5算法。C4.5算法的核心是在选择测试属性时，用信息增益率来替代信息增益，作为属性选取的标准。这一改进措施有效地克服了ID3算法具有的偏袒属性问题。在决策树构造过程中，C4.5算法引入了剪枝的策略，以此来避免数据过拟合。此外，C4.5算法增加了对连续属性离散化的处理，使得算法能够处理连续属性值。然而，算法在处理连续属性值时，需要对数据进行扫描和排序，影响了算法执行效率，且算法只能对内存中的数据进行处理。

另一个经典的决策树算法是CART算法[76]。CART算法采用代表数据不纯度的GINI系数作为属性划分的标准。相较于ID3算法和C4.5算法基于信息熵来计算测试属性，基于GINI系数的方法计算更加简便，且具有很好的近似精度。此外，CART算法用二分递归的方法进一步简化了GINI系数的计算，并得到更加简单直观的二叉决策树模型。CART算法采用类似C4.5算法的思想将连续属性离散化，因而能够处理连续属性值。不过，当属性类别过多、决策树的复杂度较高时，CART算法的误差较大。

3.3 神经网络算法

神经网络算法的核心是训练神经网络模型，即根据训练数据调整神经网络模型的参数，以使模型的表征能力达到最优。最早的神经网络学习算法是感知器训练法则[77]，该法则根据训练样例的目标输出和实际输出的差值来调整网络连接权，直至感知器能正确分类所有的训练数据。感知器训练法则对于线性可分的训练数据是收敛的[78]，但对于线性不可分的训练数据来说，它将无法收敛。针对线性不可分的训练样本，Widrow B等人[79]提出了最小均方误差（least mean square，LMS）算法。算法的核心思想是最小化输出误差的平方和，从而得到最优近似解。算法采用梯度下降的搜索策略，迭代地沿误差梯度的反方向更新网络连接的权值，直至收敛到稳定状态。Werbos P[80]提出了将LMS算法推广到由非线性可微神经元组成的多层前馈神经网络的思想。Rumelhart D E等人[81]提出了基于同样思想的误差反向传播（error back propagation，BP）算法。BP算法是目前应用最为广泛、最具代表性的神经网络学习算法。除了前馈神经网络模型，径向基函数（radical basis function，RBF）神经网络、递归神经网络、卷积神经网络等大部分神经网络模型也可以采用BP算法进行训练。

尽管BP算法具有强大的学习和表征能力及广泛的适用性，但标准的BP算法也存在着许多缺陷和不足。首先，由于算法的学习能力过于强大，其在网络训练中会出现“过拟合”现象；其次，算法有一定概率陷入局部极小，无法收敛于全局最优解；此外，算法的收敛速度受到学习率、初始条件等多个因素的影响，收敛速度较慢。针对BP算法存在的问题，研究者探索了许多改进方法。对于“过拟合”问题，一个经典的解决方法是将数据划分为训练集和测试集，当模型在测试集合上的误差达到最小时即停止迭代[82]；Barron A R[83]提出了正则化的方法，在网络的误差目标函数中引入惩罚项，从而使网络模型尽可能精简。对于算法会收敛于局部极小的问题，采用在梯度更新法则中引入冲量、使用随机梯度下降或使用不同初始值初始化多个网络并最终合并训练结果的方法，能够在一定程度上避免陷入局部极小[82]；Aarts E H等人[84]提出用模拟退火技术来优化BP算法，使算法在每一次迭代中都以一定概率接受次优解，从而跳出局部极小；此外，利用遗传算法等进化算法[85]结合BP算法训练神经网络，也可以使网络更好地接近全局最优。对于BP算法训练速度缓慢的问题，Jacobs R A[86]提出自适应学习率的改进策略，加速神经网络的训练；引入冲量的策略也可以加快神经网络的训练速度[82]。除了BP算法外，还有诸多针对特定网络模型的算法。有代表性的算法包括自组织映射（self organizing map，SOM）算法[87]、自适应谐振理论（adaptive resonance theory，ART）网络学习算法[88]和训练受限玻尔兹曼机的CD算法[89]等。

3.4 聚类分析算法

按照算法的基本思想，聚类分析算法主要分为层次聚类、基于划分的聚类、基于密度的聚类和基于网格的聚类。

层次聚类算法的基本思想是逐层将数据分组，形成一个层级式的树状图结构的聚类结果。根据构造方式的不同，层次聚类可分为两大类：聚合层次聚类和分解层次聚类。聚合层次聚类采用自底向上的方式，初始时将每个个体看作一类，再逐层合并这些类；分解层次聚类则采用自顶向下的方式，初始时将所有个体看作一类，再逐层分割这些类。

早期的较为著名的层次聚类算法是Zhang T等人[90]提出的BIRCH算法。算法的核心思想是建立一个聚类特征树（CF-Tree），并在聚类特征树上执行聚类分析。由于只需要扫描一次数据集，且聚类过程在内存中完成，因此BIRCH算法具有很高的执行效率。不过，BIRCH对于非凸的数据集分布簇的聚类效果并不理想。为了处理海量数据，识别不同形状和大小的聚类，Guha S等人[91]提出了CURE算法。该算法使用数据空间中的多个点表示一个簇，以此过滤孤立点，并能更好地识别非球状、大小不一的簇。此外，算法采用随机抽样和分区的策略处理大规模数据，以此获得更好的时间效率。ROCK算法[92]是对CURE算法的改进。ROCK算法在CURE算法的基础上，增加了对类别属性的识别，并通过考察数据点间的相似性和共同邻居数量来提高算法的顽健性。此外，Karypis G等人[93]提出了Chameleon算法。该算法采用动态建模技术，将数据集构造成k-最近邻图，并用图划分算法将其划分为若干子图，再利用层次聚类算法合并这些子图，得到聚类结果。该算法能够处理复杂形状的簇。

基于划分的聚类算法首先需要指定一个聚类数目，算法通过迭代来逐步优化一个目标函数，最终得到指定数目的结果簇。K-means算法[94]是典型的基于划分的聚类算法。算法用每个类别中全部数据的均值，即聚类中心来表示这个类。算法从k个随机的聚类中心开始，迭代地将距离聚类中心最近的点划分为一类，直至聚类中心点达到收敛。该算法简洁高效，时间和空间复杂度低，因而在聚类分析中有着十分广泛的应用。然而，K-means算法存在许多不足。K-means算法只能处理数值型的数据，且算法对非标准正态分布和非均匀样本集的聚类效果较差；算法对初始值的设置很敏感，初始聚类中心将对聚类结果产生很大影响；此外，算法对异常值数据和离群点很敏感。

为了克服K-means算法的不足，研究者对其进行了诸多改进。针对K-means算法只能处理数值型数据的问题，Huang Z[95]提出了K-modes算法，将K-means算法扩展到能够处理分类属性，并将K-means算法和K-modes算法结合，提出了能够处理混合属性的K-Prototype算法。针对K-means算法对非标准正态分布和非均匀样本处理效果差的问题，Girolami M[96]提出了基于核函数的算法，将数据映射到高维空间中进行处理；针对K-means算法对初始值敏感的问题，Arthur D等人[97]提出了改进的K-means++算法，算法的核心思想是选取k个相距较远的点作为聚类中心，K-means++算法选取聚类中心的思想形象直观，且行之有效；针对K-means算法对离群点和异常点敏感的问题，Kaufman L等人[98]提出了K-medoids算法，用簇中的某个点而非所有点的平均值来代替这个类，由此实现了对异常值的有效处理。

基于密度的聚类算法将具有一定稠密程度的数据划分为一个簇，因而能够处理任意形状的聚类，并有效排除稀疏的异常点。DBSCAN算法[99]是经典的基于密度的聚类算法。DBSCAN算法从任意未标记点出发，将密度相连的最大点集作为一个簇，并用同样的方法得到所有的聚类结果。DBSCAN算法不需要指定类别个数，能够处理任意形状的数据，对异常点不敏感。不过，由于DBSCAN使用全局密度阈值，若簇的密度分布不均匀，那么算法会将密度低于阈值的簇全部视为异常点。对此，Ankerst M等人[100]提出了OPTICS算法，将邻域点按照密度大小排序，并用可视化的方法来寻找不同密度的聚类。Ertöz L等人[101]提出SNN算法，用每对点间共享邻域的范围代替DBSCAN算法中的密度，实现对不同密度簇的聚类。

基于网格的聚类算法将数据空间划分为有限数目的网络单元，计算映射到每个单元中的数据密度，并将相邻的稠密单元合并成聚类结果。这类算法的计算时间与数据的数目和输入顺序无关，并且能够聚类各种形状的数据。不过，由于聚类的精度取决于划分的网格单元数，算法聚类质量的提高是以牺牲时间为代价的。典型的基于网格的聚类算法是STING算法[102]。算法针对不同级别的分辨率，将数据空间划分为多个层次的矩形单元，其中高层单元被划分为多个底层单元，每个单元属性的统计信息被预先计算和存储，以便执行查询操作。算法由某一层次的单元开始，逐层向下查询满足约束条件的单元，得到的查询结果等价于聚类结果。STING算法便于实现并行和增量式更新，并且具有很高的执行效率。不过，算法只能得到边界竖直或水平的聚类，聚类结果的准确性欠佳。STING+算法[103]是对STING算法的改进，用于处理动态进化的空间数据。CLIQUE算法[104]将基于密度的聚类思想与基于网格的聚类思想结合起来，因而既可以聚类任意形状的簇，又能处理高维数据。CLIQUE算法的缺点是聚类结果对于密度阈值敏感、可能受到噪声点的影响，且算法效率不高。此外，Sheikholeslami G等人[105]将信号处理的思想运用到聚类分析中，提出了WaveCluster算法。算法对网格数据结构进行小波变换，并将变换后的空间中的高密度区域识别为簇。该算法效率高，能处理高维数据，能够有效处理噪声点。

4 结束语

工业是产生大数据的重要源头，工业大数据分析有助于提高两化融合水平，实现智能制造，具有广阔的应用前景，同时也面临着诸多挑战。目前，尽管许多研究尝试将大数据分析方法应用于工业场景，但这些大数据分析模型相对简单和单一；分析算法也以通用的数据挖掘算法为主。此外，由于工业过程是高度连续化的，直接将理论研究的模型和算法部署于真实工业场景中进行实验验证，很大程度上会影响生产线的正常运行。因此，如何平衡实验的真实性和生产过程的稳定性，也是进行工业大数据分析的研究需要面临的关键问题。未来，工业大数据分析模型和算法的研究可以从以下几个方面展开。

- 设计表征能力更强、结构更复杂的模型来描述复杂工业过程。
- 设计或改进现有的大数据分析方法，使其能够满足实际工业场景对准确性、实时性、可扩展性的要求。
- 设计模拟器，依据工业生产的规律

模拟实际生产过程并产生数据，为工业大数据分析模型和算法的研究提供仿真实验环境。

参考文献:

[1] GE. The case for an industrial big data platform: laying the groundwork for the new industrial age[R]. 2013.

[2] ZHENG L, ZENG C, LI L, et al. Applying data mining techniques to address critical process optimization needs in advanced manufacturing[C]// The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, 2014, New York, USA. New York: ACM Press, 2014: 1739-1748.

[3] BERKHIN P. A survey of clustering data mining techniques[J]. Grouping Multidimensional Data, 2006, 43(1): 25-71.

[4] HIPP J, NAKHAEIZADEH G. Algorithms for association rule mining: a general survey and comparison[J]. ACM SIGKDD Explorations Newsletter, 2000, 2(1): 58-64.

[5] COOLEY R, MOBASHER B, SRIVASTAVA J. Web mining: information and pattern discovery on the world wide web[C]//The 9th International Conference on Tools with Artificial Intelligence, November 3-8, 1997, Newport Beach, USA. Washington, DC: IEEE Computer Society, 1997: 558.

[6] 任磊, 杜一, 马帅, 等. 大数据可视分析综述[J]. 软件学报, 2014 (9): 1909-1936.
REN L, DU Y, MA S, et al. Visual analytics towards big data[J]. Journal of Software, 2014(9): 1909-1936.

[7] STEYERBERG E W, JR H F, BORSBOOM G J, et al. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis[J]. Journal of Clinical Epidemiology, 2001, 54(8): 774-781.

[8] VYAS R, SHARMA L K, VYAS O P, et al. Associative classifiers for predictive analytics: comparative performance study[M]. Piscataway: IEEE Press, 2008: 289-294.

[9] HAAS P J, MAGLIO P P, SELINGER P G, et al. Data is dead... without what-if models[J]. PVLDB, 2012, 4(4): 1486-1489.

[10] EVANS J R, LINDNER C H. Business analytics: the next frontier for decision sciences[J]. Decision Line, 2012, 43(2): 4.

[11] 崔妍, 包志强. 关联规则挖掘综述[J]. 计算机应用研究, 2016, 33(2): 330-334.
CUI Y, BAO Z Q. Survey of association rule mining[J]. Application Research of Computers, 2016, 33(2): 330-334.

[12] LIU C Y, SUN Y F. Application of data mining in production quality management[J]. International Symposium on Intelligent Information Technology Application, 2009(2): 284-287.

[13] 娄小芳. 基于模式识别和数据挖掘的铝工业生产节能降耗研究[D]. 长沙: 国防科学技术大学, 2010.
LOU X F. Research on energy saving of the aluminum industrial production based on technologies of pattern recognition and data mining[D]. Changsha: National University of Defense Technology, 2010.

[14] 张健. 服装制造数据协同与辅助决策系统的研究与设计[D]. 苏州: 苏州大学, 2013.
ZHANG J. Research and design of the garment manufacturing data collaboration and auxiliary decision-making system[D]. Suzhou: Soochow University, 2013.

[15] 薛百里. 基于数据挖掘技术的服装制造标准工时制定方法研究[D]. 苏州: 苏州大学, 2015.
XUE B L. Research on garment manufacturing standard working hour formulation method based on data mining[D]. Suzhou: Soochow University, 2015.

[16] 谢英星. Apriori算法在模具改模工艺信息

处理中的应用[J]. 组合机床与自动化加工技术, 2008(6): 80-83.
XIE Y X. The application of Apriori algorithm in mould repair process management[J]. Modular Machine Tool & Automatic Manufacturing Technique, 2008(6): 80-83.

[17] AGARD B. Data mining for improvement of product quality[J]. International Journal of Production Research, 2006, 44(18-19): 4027-4041.

[18] 张亮. 数据挖掘在机械制造业外购件供应系统的应用[D]. 重庆: 重庆大学, 2004.
ZHANG L. Realization of data mining on machinery enterprise marketing system[D]. Chongqing: Chongqing University, 2004.

[19] 周明. 基于数据挖掘的制造业采购DSS理论及方法研究[D]. 天津: 天津大学, 2009.
ZHOU M. Study on theory and method of manufacturing procurement dss based on data mining[D]. Tianjin: Tianjin University, 2009.

[20] 王建良, 杜元胜, 徐建良. 面向离散制造业数据挖掘技术研究与应用[J]. 微计算机信息, 2007(33): 10-11, 21.
WANG J L, DU Y S, XU J L. Research and application of data mining in intermittent manufacturing industry[J]. Control & Automation, 2007(33): 10-11, 21.

[21] 李烨. 数据挖掘技术在卡车制造商的客户价值分析应用研究[D]. 柳州: 广西科技大学, 2013.
LI Y. The research for the application of data mining technology in the customer value analysis of the trunk manufactures[D]. Liuzhou: Guangxi University of Technology, 2013.

[22] 汪奇, 黄洪, 郑晓群. 笔记本电脑BTO生产计划中关联规则挖掘的应用研究[J]. 商场现代化, 2007(36): 30-31.
WANG Q, HUANG H, ZHENG X Q. The application research of association rules mining in BTO production plan of laptop[J]. Market Modernization, 2007(36): 30-31.

[23] 黄亦弢 钟表供应链管理中智能物料表研究[D]. 广州: 广东工业大学, 2006.
HUANG Y T. Research on intelligent bill of material in the supply chain management of horologe industry[D]. Guangzhou: Guangdong University of Technology, 2006.

[24] 石慧. 基于汽车服务业的服务挖掘[J]. 上海汽车, 2007(7): 28-30.
SHI H. Service "mining" based on automobile service industry[J]. Shanghai Auto, 2007(7): 28-30.

[25] 陈嵩. 通信设备制造业CRM中的数据挖掘研究[D]. 武汉: 武汉理工大学, 2008.
CHEN S. Data mining research in communication equipment manufacturing industry CRM[D]. Wuhan: Wuhan University of Technology, 2008.

[26] CHEN W C, TSENG S S, WANG C Y. A novel manufacturing defect detection method using data mining approach[C]// Innovations in Applied Artificial Intelligence, International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, IEA/AIE 2004, May 17-20, 2004, Ottawa, Canada. Heidelberg: Springer, 2004: 77-86.

[27] IFLIKLI C, KAHYA, ZYIRMIDOKUZ E. Implementing a data mining solution for enhancing carpet manufacturing productivity[J]. Knowledge-Based Systems, 2010, 23(8): 783-788.

[28] LI M, FENG S, SETHI I K, et al. Mining production data with neural network&CART[C]//The 3rd IEEE International Conference on Data Mining, November 19-22, Melbourne, USA. Piscataway: IEEE Press, 2003: 783-788.

[29] 宋旭东, 刘晓冰, 程晓兰, 等. 钢铁企业生产成本关键工序数据挖掘应用研究[J]. 计算机工程与应用, 2008(28): 184-186, 195.
SONG X D, LIU X B, CHENG X L, et al. Research on production cost key processes data mining for iron & steel enterprises[J].

Computer Engineering and Applications, 2008(28): 184-186, 195.

[30] 郭龙波. 基于数据挖掘方法的冷轧表面质量缺陷分析[D]. 马鞍山: 安徽工业大学, 2012.
GUO L B. The defect analysis for the surfacial quality of the cooling system based on data mining technology[D]. Ma'anshan: Anhui University of Technology, 2012.

[31] 王诗. 基于数据挖掘技术的矿用提升机故障预警系统的研究[D]. 北京: 北京邮电大学, 2009.
WANG S. Research on fault diagnosis and safety alarm[D]. Beijing: Beijing University of Posts and Telecommunications, 2009.

[32] 王成龙. 基于数据挖掘技术的生产调度问题研究[D]. 杭州: 浙江大学, 2015.
WANG C L. Research on data mining based production scheduling[D]. Hangzhou: Zhejiang University, 2015.

[33] 徐玉婷. MES车间生产调度系统及其数据挖掘方法的研究[D]. 南京: 南京航空航天大学, 2007.
XU Y T. Research on MES workshop scheduling system and its data mining method[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2007.

[34] 刘娜. 纺织领域中数据挖掘技术的应用研究[D]. 大连: 大连海事大学, 2004.
LIU N. Research on the application of data mining technology in textile area[D]. Dalian: Dalian Maritime University, 2004.

[35] 郭玲. 信息时代汽车制造业协同采购策略研究[D]. 长春: 吉林大学, 2006.
GUO L. Study on strategy of collaborative procurement business in auto manufacturing in information times[D]. Changchun: Jilin University, 2006.

[36] 陈力. 大数据环境下集团企业的精细化营销体系设计方法与实现[D]. 杭州: 浙江理工大学, 2016.
CHEN L. The implementation and design method of precise marketing system for enterprise under big data environment[D]. Hangzhou: Zhejiang Sci-Tech University, 2016.

[37] 刘菲. 基于决策树技术的忠诚客户挖掘研究[D]. 阜新: 辽宁工程技术大学, 2009.
LIU F. Research of loyal customer mining based on decision tree technology[D]. Fuxin: Liaoning Technical University, 2009.

[38] 沈小淦. 制造业售后服务系统的研究与开发[D]. 西安: 西安电子科技大学, 2012.
SHEN X G. Research and development of manufacturing after service system[D]. Xi'an: Xidian University, 2012.

[39] 陈思行, 陈保钢. 基于制造业的CRM的客户流失分析[J]. 建设机械技术与管理, 2008(3): 106-109.
CHEN S X, CHEN B G. Analysis on the loss of clients based on manufacturing industry CRM[J]. Construction Machinery Technology & Management, 2008(3): 106-109.

[40] 鲁钊, 陈世平. 基于ID3算法的机械制造业决策应用[J]. 计算机应用, 2011(11): 3087-3090.
LU Z, CHEN S P. Application of machinery manufacturing decision-making based on ID3 algorithm[J]. Journal of Computer Applications, 2011(11): 3087-3090.

[41] 骆自超. 基于数据挖掘的发动机缸盖燃烧室容积制造误差控制方法研究[D]. 上海: 上海交通大学, 2014.
LUO Z C. The construction of optimal control scheme of volume of automotive cylinder head combustion chamber based on data mining techniques[D]. Shanghai: Shanghai Jiao Tong University, 2014.

[42] MOGHIMI M, SARAEE M H, BAGHERI A. Modeling of batch annealing process using data mining techniques for cold rolled steel sheets[C]//The 1st International Workshop on Data Mining for Service and Maintenance, August 21, 2011, San Diego, USA. New York: ACM Press, 2011: 18-22.

[43] SHI X, SCHILLINGS P, BOYD D. Applying

artificial neural networks and virtual experimental design to quality improvement of two industrial processes[J]. International Journal of Production Research, 2004, 42(1): 101-118.

[44] 刘立强. 企业能源管理系统的数据整合与处理[D]. 北京: 北京交通大学, 2014.
LIU L Q. Data integration and processing in the enterprises energy management system[D]. Beijing: Beijing Jiaotong University, 2014.

[45] 吴波. 造纸过程能源管理系统中数据挖掘与能耗预测方法的研究[D]. 广州: 华南理工大学, 2012.
WU B. Study on data mining and prediction method of energy consumption used in energy management system of paper process[D]. Guangzhou: South China University of Technology, 2012.

[46] HONG S J, LIM W Y, CHEONG T, et al. Fault detection and classification in plasma etch equipment for semiconductor manufacturing e-diagnostics[J]. IEEE Transactions on Semiconductor Manufacturing, 2012, 25(1): 83-93.

[47] 罗洪波. 汽车售后服务故障件管理及数据挖掘技术应用研究[D]. 成都: 西南交通大学, 2008.
LUO H B. The invalid parts management of automobile in after-sales service and application research on data mining[D]. Chengdu: Southwest Jiaotong University, 2008.

[48] 范卿. 工程机械远程监控系统研究[D]. 长沙: 湖南大学, 2011.
FAN Q. Research on remote monitoring system for construction machinery[D]. Changsha: Hunan University, 2011.

[49] 于欣. 基于数据挖掘的物流设备隐性故障预警模型研究[D]. 秦皇岛: 燕山大学, 2014.
YU X. Research of logistics equipment hidden failure warning model based on data mining[D]. Qinhuangdao: Yanshan University, 2014.

[50] 孙宜然, 赵嵩正, 徐伟. 面向供应链的制造业库存决策支持系统的分析与设计[J]. 工业工程, 2006(3): 75-79.
SUN Y R, ZHAO A Z, XU W. Supply chain oriented decision-making support system for manufacturing inventory management[J]. Industrial Engineering Journal, 2006(3): 75-79.

[51] 陈承贵. 基于数据挖掘技术仓库管理系统的应用与研究[D]. 成都: 电子科技大学, 2009.
CHEN C G. Application and research of warehouse management system based on data mining technology[D]. Chengdu: University of Electronic Science and Technology of China, 2009.

[52] 文小敏, 袁清珂, 罗小美. ERP决策支持系统中采购数据仓库的构建与应用研究[J]. 现代机械, 2005(2): 27-29.
WEN X M, YUAN Q H, LUO X M. The building and application of purchase data warehouse in DSS of ERP[J]. Modern Machinery, 2005(2): 27-29.

[53] 桂卫华, 黄泰松, 朱爽. 智能综合原料库存优化系统及应用[J]. 中南工业大学学报(自然科学版), 2001(5): 536-540.
GUI W H, HUANG T S, ZHU S. Intelligent integrated raw material storage optimization system and application[J]. Journal of Central South University of Technology(Natural Science), 2001(5): 536-540.

[54] 周志刚. 灰色系统理论与人工神经网络融合的时序数据挖掘预测技术及应用[D]. 成都: 成都理工大学, 2006.
ZHOU Z G. The fusion technology between grey system theories and neural networks and its application in prediction for time sequence[D]. Chengdu: Chengdu University of Technology, 2006.

[55] 王鹏鹏, 廖小平, 邓建新. ERP环境下销售决策支持的研究与实现[J]. 机械设计与制造, 2008(5): 219-221.
WANG P P, LIAO X P, DENG J X. Research and implementation of sales decision support in ERP environment[J]. Machinery Design & Manufacture, 2008(5): 219-221.

[56] 李霄林. 面向摩托车智能设计的数据挖掘系统研究与应用[D]. 重庆: 重庆大学, 2006.
LI X L. The research and application for the data mining system of motorcycle intelligent design[D]. Chongqing: Chongqing University, 2006.

[57] AYE T T, YANG F, WANG L, et al. Data driven framework for degraded pogo pin detection in semiconductor manufacturing[C]//The 10th IEEE Conference on Industrial Electronics and Applications, June 15-17, 2015, Oakland, NowZealand. Pisoataway: IEEE Press, 2015.

[58] 常建涛, 仇原鹰, 李申, 等. 生产计划与调度中的次年产量预测方法[J]. 计算机集成制造系统, 2013(7): 1648-1654.
CHANG J T, QIU Y Y, LI S, et al. Output prediction approach of production planning and scheduling in the next year[J]. Computer Integrated Manufacturing Systems, 2013(7): 1648-1654.

[59] 梁琦, 陆剑宝. 传统制造业集群的生产性服务需求——广东、山西两地4个制造业集群样本的考察[J]. 管理评论, 2014(11): 169-181.
LIANG Q, LU J B. The producer services demand of traditional manufacturing clusters: an investigate of four manufacturing clusters samples in Guangdong and Shanxi[J]. Management Review, 2014(11): 169-181.

[60] 江小辉, 赵建民, 朱信忠. 商业智能在流行饰品制造业中的应用研究[J]. 信息技术, 2008(6): 33-35.
JIANG X H, ZHAO J M, ZHU X Z. Application research of business intelligence in the fashion jewelry manufacture industry[J]. Information Technology, 2008(6): 33-35.

[61] 张玉东. PG炼钢厂MES系统数据挖掘的设计与开发[D]. 成都: 电子科技大学, 2011.
ZHANG Y D. Design and development of data mining for MES system of PG steelmaking plant[D]. Chengdu: University of Electronic Science and Technology of China, 2011.

[62] 肖溱鸽. 基于数据分析的数控加工工艺参数能效优化方法研究[D]. 重庆: 重庆大学, 2016.
XIAO Z G. Research on data analysis based process parameters optimization method for energy efficiency in CNC machining[D]. Chongqing: Chongqing University, 2016.

[63] URTUBIA A, PÉREZ-CORREA J R, SOTO A, et al. Using data mining techniques to predict industrial wine problem fermentations[J]. Food Control, 2007, 18(12): 1512-1517.

[64] 宋辉. 聚类分析系统的设计与实现及在工业中的应用[D]. 天津: 天津科技大学, 2004.
SONG H. Design and implement of clustering analysis system and its application in industry[D]. Tianjin: Tianjin University of Science&Technology, 2004.

[65] 武霞. Hadoop平台下基于聚类和关联规则算法的工程车辆故障预测研究[D]. 太原: 太原科技大学, 2015.
WU X. Research on fault prediction of engineering vehicles based on clustering and association rules algorithm under Hadoop platform[D]. Taiyuan: Taiyuan University of Science and Technology, 2015.

[66] 阮志林. 基于大批量定制模式的管理信息系统的研究及其应用[D]. 长沙: 国防科学技术大学, 2005.
RUAN Z L. Research & application of management information system based on mass customization[D]. Changsha: National University of Defense Technology, 2005.

[67] AGRAWAL R S R. Fast algorithm for mining association rules[C]//The 20th International Conference on Very Large Databases (VLDB), September 12-15, 1994, Santiago de Chile, Chile. Chile: Morgan Kaufmann Publishers Inc, 1994: 487-499.

[68] MANNILA H, TOIVONEN H, VERKAMO A I. Efficient algorithms

for discovering association rules[C]// The 3rd International Conference on Knowledge Discovery and Data Mining, July 31–August 1, 1994, Seattle, USA. Palo Alto: AAAI Press, 1994: 181–192.

[69] PARK J S, CHEN M S, YU P S. An effective hash-based algorithm for mining association rules[C]//The 1995 ACM SIGMOD International Conference on Management of Data, May 22–25, 1995, San Jose, USA. New York: ACM Press, 1995: 175–186.

[70] SAVASERE A, OMIECINSKI E R, NAVATHE S B. An efficient algorithm for mining association rules in large databases[C]//The 21th International Conference on Very Large Data Bases, September 11–15, 1995, Zurich, Switzerland. San Francisco: Morgan Kaufmann Publishers Inc., 1995: 432–444.

[71] TOIVONEN H. Sampling large databases for association rules[C]//The 22th International Conference on Very Large Data Bases, September 3–6, 1996, Mumbai, India. San Francisco: Morgan Kaufmann Publishers Inc., 1996: 134–145.

[72] AGRAWAL R, SHAFER J C. Parallel mining of association rules[J]. IEEE Transactions on Knowledge and Data Engineering, 1996, 8(6): 962–969.

[73] HAN J, PEI J, YIN Y. Mining frequent patterns without candidate generation[C]//The 2000 ACM SIGMOD International Conference on Management of Data, May 15–18, 2000, Dallas, USA. New York: ACM Press, 2000: 1–12.

[74] QUINLAN J R. Induction of decision trees[J]. Machine learning, 1986, 1(1): 81–106.

[75] QUINLAN J R. C4. 5: programs for machine learning[M]. San Francisco: Morgan Kaufmann, 1993.

[76] BREIMAN L, FRIEDMAN J, STONE C J, et al. Classification and regression trees[M]. Boca Raton: CRC Press, 1984.

[77] ROSENBLATT F. The perceptron: a probabilistic model for information storage and organization in the brain[J]. Psychological Review, 1958, 65(6): 386–408.

[78] ROSENBLATT F. Principles of neurodynamics: perceptrons and the theory of brain mechanisms[M]. New York: Spartan Books, 1962: 705.

[79] WIDROW B, HOFF M E. Adaptive switching circuits. 1960 IRE WESCON Convention Record[R]. 1960.

[80] WERBOS P. Beyond regression: new tools for prediction and analysis in the behavioral science[J]. Ph.D.Dissertation Harvard University, 1974, 29(18): 65–78.

[81] RUMELHART D E, MCCLELLAND J L, GROUP C P. Parallel distributed processing: explorations in the microstructure of cognition, vol. 2: psychological and biological models[J]. Language, 1986, 63(4): 45–76.

[82] MITCHELL T M. Machine learning[M]. Burr Ridge, IL: McGraw Hill, 1997: 870–877.

[83] BARRON A R. Complexity regularization with application to artificial neural networks[J]. Nonparametric Functional Estimation and Related Topics, 1991, 5(4): 561–576.

[84] AARTS E H, KORST J. Simulated annealing and Boltzmann machines[M]// A stochastic approach to combinatorial optimization. New York: Wiley, 1989.

[85] YAO X. Evolving artificial neural networks[J]. Proceedings of the IEEE, 1999, 87(9): 1423–1447.

[86] JACOBS R A. Increased rates of convergence through learning rate adaptation[J]. Neural Networks, 1987, 1(4): 295–307.

[87] KOHONEN T. Self-organized formation of topologically correct feature maps[J]. Biological Cybernetics, 1982, 43(1): 59–69.

[88] CARPENTER G A, GROSSBERG S. A massively parallel architecture for a self-organizing neural pattern recognition machine[J]. Computer Vision,

Graphics, and Image Processing, 1987, 37(1): 54-115.

[89] HINTON G E. A practical guide to training restricted Boltzmann machines[M]. Heidelberg: Springer, 2010: 599-619.

[90] ZHANG T, RAMAKRISHNAN R, LIVNY M. BIRCH: an efficient data clustering method for very large databases[C]// ACM SIGMOD International Conference on Management of Data, 1996: 103-114.

[91] GUHA S, RASTOGI R, SHIM K. CURE: an efficient clustering algorithm for large databases[C]//The 1998 ACM SIGMOD International Conference on Management of Data, June 1-4, 1998, Seattle, USA. New York: Acm Press 1998: 73-84.

[92] GUHA S, RASTOGI R, SHIM K. ROCK: a robust clustering algorithm for categorical attributes[C]//The 15th International Conference on Data Engineering, March 23-26, 1999, Sydney, Australia. Piscataway: IEEE Press, 1999: 512-521.

[93] KARYPIS G, HAN E-H, KUMAR V. Chameleon: hierarchical clustering using dynamic modeling[J]. Computer, 1999, 32(8): 68-75.

[94] FORGY E W. Cluster analysis of multivariate data: efficiency versus interpretability of classifications[J]. Biometrics, 1965, 21(3): 768-769.

[95] HUANG Z. Extensions to the k-means algorithm for clustering large data sets with categorical values[J]. Data Mining & Knowledge Discovery, 1998, 2(3): 283-304.

[96] GIROLAMI M. Mercer kernel-based clustering in feature space[J]. IEEE Transactions on Neural Networks, 2002, 13(3): 780-784.

[97] ARTHUR D, VASSILVITSKII S. k-means++:the advantages of careful seeding[C]//The 18th Acm-Siam Symposium on Discrete Algorithms, January 7-9, 2007, New Orleans, Louisiana. Philadelphia: Society for Industrial and Applied Mathematics, 2007: 1027-1035.

[98] KAUFMAN L, ROUSSEEUW P. Clustering by means of medoids[M]. [S.l.]: North-Holland, 1987.

[99] ESTER M, KRIEGEL H P, XU X. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise[C]//The 2nd International Conference on Knowledge Discovery and Data Mining, August 2-4, 1996, Portland, USA. Palo Alto: AAAI Press, 1996: 226-231.

[100] ANKERST M, BREUNIG M M, KRIEGEL H P. OPTICS: ordering points to identify the clustering structure[J]. ACM SIGMOD Record, 1999, 28(2): 49-60.

[101] ERTÖZ L, STEINBACH M, KUMAR V. A new shared nearest neighbor clustering algorithm and its applications[C]// The Workshop on Clustering High Dimensional Data & ITS Applications at Siam International Conference on Data Mining, April 11-13, 2002, Arlington, Virginia. [S.l.:s.n.], 2002.

[102] WANG W, YANG J, MUNTZ R R. STING: a statistical information grid approach to spatial data mining[C]// The 23rd International Conference on Very Large Data Bases, August 25-29, 1997, Athens, Greece. San Francisco: Morgan Kaufmann Publishers Inc., 1997: 186-195.

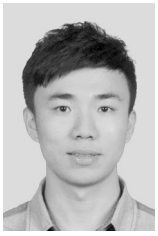
[103] WANG W, YANG J, MUNTZ R. STING+: an approach to active spatial data mining[C]//The 15th International Conference on Data Engineering, March 23-26, 1999, Sydney, Australia. Washington, DC: IEEE Computer Society, 1999: 116-125.

[104] AGRAWAL R, GEHRKE J, GUNOPULOS D, et al. Automatic subspace clustering of high dimensional data for data mining applications[M]. New York: ACM Press, 1998: 94-105.

[105] SHEIKHOLESLAMI G, CHATTERJEE S, ZHANG A. Wavecluster: a multi-resolution clustering approach for very large spatial databases[C]//The 24rd International Conference on Very Large Data Bases, August 24-27, 1998, New York, USA. San Francisco: Margan Kaufmann, 1998: 428-439.

作者简介

王宏志(1978-),男,博士,哈尔滨工业大学计算机科学与技术学院教授,博士生导师,主要研究方向为大数据。



梁志宇(1994-),男,哈尔滨工业大学计算机科学与技术学院硕士生,主要研究方向为大数据。

李建中(1950-),男,哈尔滨工业大学计算机科学与技术学院教授,博士生导师,主要研究方向为大数据、物联网。

高宏(1966-),女,博士,哈尔滨工业大学计算机科学与技术学院教授,博士生导师,主要研究方向为大数据、物联网。

收稿日期: 2018-07-18

基金项目: 国家自然科学基金资助项目(No.U1509216, No.61472099, No.61602129);国家重点研发计划基金资助项目(No. 2016YFB1000703);黑龙江省留学回国人员基金资助项目(No.LC2016026)

Foundation Items: The National Natural Science Foundation of China(No.U1509216, No.61472099, No.61602129), The National Key Research and Development Program of China(No.2016YFB1000703), The Scientific Research Foundation for the Returned Overseas Chinese Scholars of Heilongjiang Provience(No.LC2016026)