# 制造业生产过程中多源异构数据处理方法综述

**陈世超[1,2]，崔春雨[1]，张华[3]，马戈[4]，朱凤华[1]，商秀芹[1]，熊刚[1]**
1. 中国科学院自动化研究所复杂系统管理与控制国家重点实验室，北京 100190；
2. 澳门科技大学，澳门 999078；3. 北京航天智造科技发展有限公司，北京 100039；
4. 中国工业互联网研究院，北京 100102

**摘要**
随着现代制造业向着自动化、信息化、智能化方向快速发展，生产过程中会产生大量的多源异构数据。对多源异构数据的有效处理和深度挖掘可为生产制造者提供更有效的生产调度、设备管理等策略，从而提高生产质量和效率。针对制造业生产过程中多源异构数据的处理方法与技术等进行系统性的综述，首先明确了制造业生产过程多源异构数据内容及分类；其次，阐述了多源异构数据处理中数据采集、数据集成及数据分析各个阶段应用的数据处理方法和技术，并分析了各种方法与技术的优缺点以及应用；最后，对生产过程中多源异构数据处理方法和技术进行总结，指出了现阶段多源异构数据处理方法及技术面临的挑战和发展趋势。

**关键词**
数据处理；多源异构数据；生产制造

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2020044

# *A survey on multi-source heterogeneous data processing methods in manufacturing process*

**CHEN Shichao[1,2], CUI Chunyu[1], ZHANG Hua[3], MA Ge[4], ZHU Fenghua[1], SHANG Xiuqin[1], XIONG Gang[1]**
1. The State Key Laboratory for Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
2. Macau University of Science and Technology, Macau 999078, China
3. Beijing Aerospace Smart Manufacturing Technology Development, Beijing 100039, China
4. China Academy of Industrial Internet, Beijing 100102, China

***Abstract***
The effective processing and deep mining analysis of data can provide manufacturers with more effective production scheduling, equipment management and other strategies to affect production yield and efficiency. The processing methods and technologies of multi-source heterogeneous data in the manufacturing process were systematically reviewed. Firstly, the multi-source heterogeneous data content and classification in the manufacturing process was clarified. Secondly, the data processing methods and techniques applied in various stages of data collection, data integration and data analysis in the multi-source heterogeneous data processing were described. And the advantages and disadvantages of various techniques and their applications were analyzed. Finally, the multi-source heterogeneous data processing methods and techniques in the manufacturing process were summarized. And the challenges and development trends were pointed out.

***Key words***
data processing, multi-source heterogeneous data, manufacturing

## 1 引言

在全球信息技术快速发展的背景下，随着科学技术的迅猛发展和社会信息化程度的不断提高，人类社会共享的数据的数量大大增加，共享的数据的形式大大丰富。据希捷公司与国际数据公司（IDC）共同发布的《数字化世界——从边缘到核心》白皮书，全球数据圈规模将从2018年的33 ZB增至2025年的175 ZB。其中，白皮书中指出，在全球数据圈中，制造业数据所占份额最大，远远超过其他行业。同时，伴随着中国“智能制造 2025”国家战略的实施，工业制造业面临重大的变革转型，大数据成为提升制造业生产力、创造力的关键。随着智能制造的发展，自动化、信息化、智能化等技术渗透到制造业生产过程的各个环节，从工业现场的传感器、设备到制造生产过程中的各个信息系统（如制造执行管理系统、生产监控系统、设备运行维护系统、产品质量检测系统、能耗管理系统等），均会产生大量不同结构类型的数据。以一个典型的纺织制造车间为例，其一天的数据量将达到84 GB[1]，而一台半导体生产机器一天的数据量甚至可以达到TB级别，这些数据包括二进制、文本、视频、音频等数据。而海量的数据中蕴含着大量有价值的信息，对这些信息的提取有利于指导人们在生产制造、设备管理和生产调度等过程中做出正确的决策，达到优化制造流程、提高效能的目的，促进制造业生产过程的全面智能化，从而提高生产质量和效率。

如**图1**所示，产品的制造流程包括研发设计、物料采购、生产制造、产品销售及产品售后5个阶段，每个阶段的数据都具有数据来源多样、数据质量低、数据蕴含信息复杂、数据实时性高等特点，而从海量数据中发掘指导制造业研发设计、生产制造、销售售后和经营管理等过程的知识和规则，需要大量的模型算法等数据处理方法的支撑。尤其是在产品生产制造过程中产生的数据，其不仅数据量十分庞大，来源丰富、类型多样、结构复杂，而且由于制造业不同的部门和系统之间数据的来源、存储形式等各不相同，数据源之间存在异构性、分布性和自治性，数据类型既包括数字、关系型数据等结构化数据，也包括图像、音频等非结构化数据。因此，这对制造业生产制造过程中海量数据的处理方法和技术提出了更高的要求。为了充分发挥制造业多源异构数据信息的潜力，更加高效地进行数据处理，必须在明确多源异构数据概念的基础上，对多源异构数据的处理方法和技术展开深入且系统性的研究。

本文首先明确了制造业生产过程中多源异构数据的概念和类型；其次对生产过程中多源异构数据处理的过程进行了划分，同时对各个阶段的数据处理方法和技术及其在制造业生产过程中的应用进行了深入分析与讨论；最后，对生产过程中多源异构数据处理方法及技术进行了总结，并对现阶段面临的挑战及未来的发展趋势进行了分析与讨论。

## 2 制造业生产过程中的多源异构数据

《大数据：下一个创新、竞争和生产力的前沿》[2]针对社会对大数据的关注及应用需求，对海量数据的处理技术进行了介绍和总结。基于对不同来源、多种结构数据的综合研究的迫切需要[3]，多源异构数据这一概念随之产生，其主要包括两个特征：一是数据来源具有多源性；二是数据种类及形态具有复杂性，即异构性[4]。

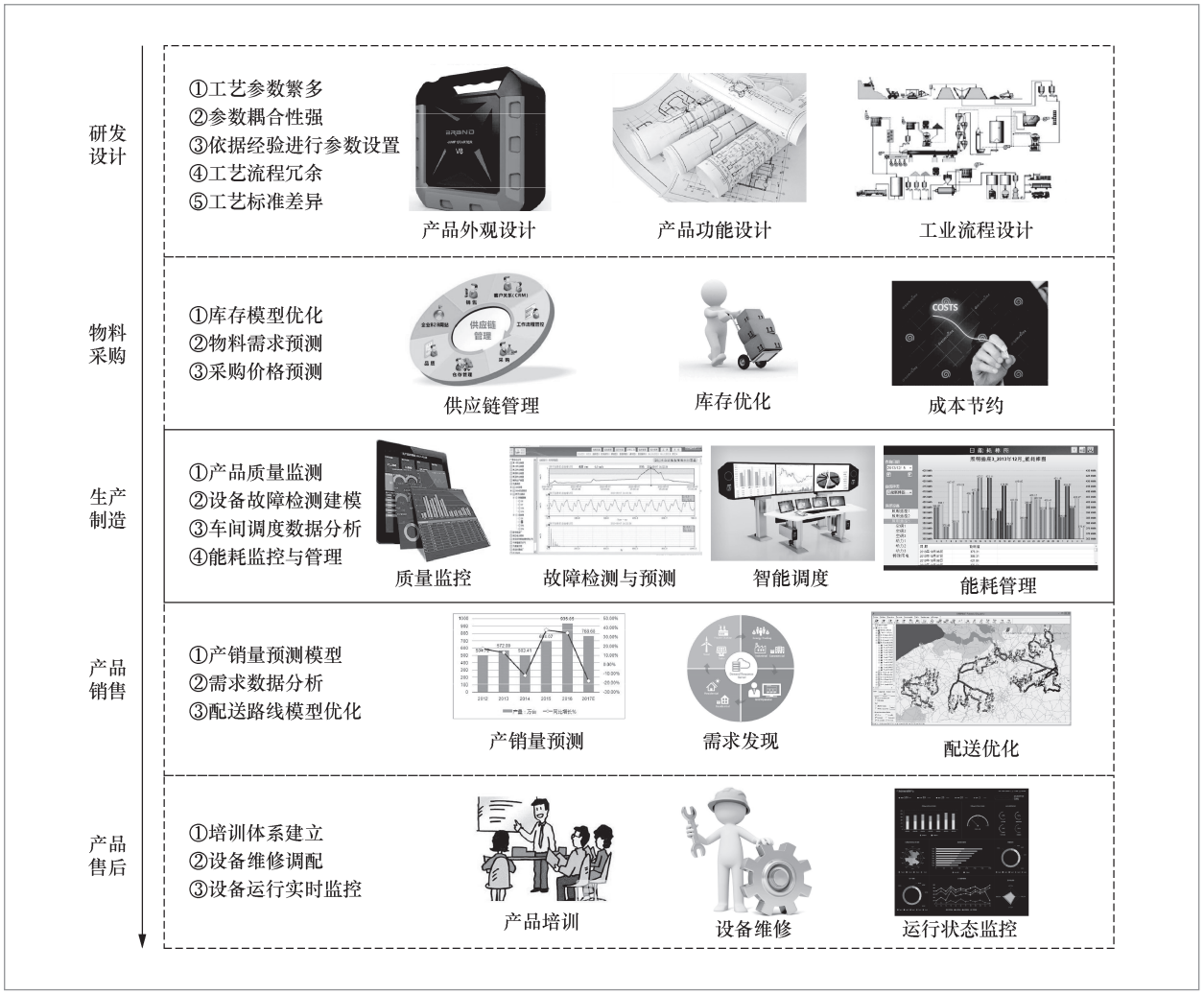


图1 制造流程的5个阶段

多源异构数据来自多个数据源，包括不同数据库系统和不同设备在工作中采集的数据集等[5]。不同的数据源所在的操作系统、管理系统不同，数据的存储模式和逻辑结构不同，数据的产生时间、使用场所、代码协议等也不同，这造成了数据“多源”的特征[6]。

另外，多源异构数据包括多种类型的结构化数据、半结构化数据和非结构化数据。结构化数据指关系模型数据，即以关系数据库表形式管理的数据；半结构化数据指非关系模型的、有基本固定结构模式的数据，例如日志文件、XML文档、JSON文档、E-mail等；非结构化数据指没有固定模式的数据，如WORD、PDF、PPT、EXL及各种格式的图片、视频等。不同类型的数据在形成过程中没有统一的标准，因此造成了数据“异构”的特征。

随着自动化、信息化、智能化等技术在制造业中的广泛应用，在生产过程中必然会产生大量的多源异构数据。从数据的来源来说，制造业的制造执行管理系统、生产监控系统、设备运行维护系统、产品质量检测系统、能耗管理系统中的各种机器

设施、工业传感器等在运行和维护过程中都会产生大量的数据。从数据结构类型来看，这些海量多源异构数据既包括设备监测数据、产品质量检测数据、能耗数据等结构化数据，还包括生产监控系统产生的大量图片、视频等非结构化数据[7]。本文综合其他学者的研究基础，针对制造业生产过程中产生的数据，按照数据来源和类型，将其做如下划分，见**表1**。对于制造业生产过程中的多源异构数据来说，由于生产过程存在复杂的变化条件，因此对数据的全面性、实时性的要求较高[8]。

## 3 制造业生产过程中多源异构数据处理

在制造业生产过程中，从前期的数据广泛采集，到最后数据的价值提取，多源异构数据处理的一般流程包括数据采集、数据集成及数据分析。数据采集主要实现大量原始数据准确、实时的采集，为数据集成阶段提供原始数据源。数据集成主要实现数据的数据库存储，数据清洗、转换、降维等预处理以及构建海量关联数据库，为数据分析阶段提供预处理的数据源。数据分析主要利用关联分析、分类聚类及深度学习等技术实现数据的价值挖掘。多源异构数据处理的一般流程如**图2**所示[14]。

### 3.1 数据采集

数据采集是多源异构数据处理的基础，只有实现对生产过程中产生的大量原始数据准确、实时的采集，并将其传输到数据存储管理平台，才能对生产设备、产品质量、工作调度等进行监控与管理，从而帮助生产管理部门做出更高效、精准的决策。

针对不同类型生产制造业生产过程中的多源异构数据，需要采用不同的数据采集方法和工具。首先，对于离散制造业中的生产过程数据，主要使用射频识别（radio frequency identification，RFID）技术[15]对生产车间中的原材料、设备、产品信息等进行数据采集。针对生产流水线上的产品信息，曹伟等人[16]提出了一种无

表1 制造业生产过程中多源异构数据划分

| 数据名称 | 数据内容[9-13] | 数据来源 | 数据类型 |
|---|---|---|---|
| 设备属性 | 生产日期、规格型号、编号、性能等 | 设备运行维护系统 | 结构化 |
| 能耗数据 | 用电量等能耗数据 | 能耗管理系统 | 结构化 |
| 生产计划 | 人员配置、排班表等 | 制造执行管理系统 | 非结构化 |
| 运行信息 | 设备温度、电流、电压等 | 生产监控系统 | 结构化 |
| 环境参数 | 光电、热敏、声敏、湿敏等工业传感器信息 | 生产监控系统 | 结构化 |
| 产品生产信息 | 产品尺寸、数量等 | 生产监控系统 | 结构化 |
| 产品质量信息 | 产品合格数、合格率等 | 产品质量检测系统 | 结构化 |
| 网络公开数据 | 电子商务网站产品报价、搜索引擎产品搜索次数等 | 公共服务网络 | 结构化 |
| 接口数据 | 接口类型数据（JSON格式、XML格式） | 已建成的工业自动化或信息系统 | 半结构化 |
| 物料数据 | 生产原料相关图文数据信息等 | 生产供应系统 | 非结构化 |
| 知识数据 | 专利、专著、企业文献等 | 制造执行管理系统 | 非结构化 |
| 产品文档 | 工程图纸、仿真数据、测试数据等 | 制造执行管理系统 | 非结构化 |
| 生产监控图片 | 图像设备拍摄的图片 | 生产监控系统 | 非结构化 |
| 生产监控音频 | 语音及声音信息 | 生产监控系统 | 非结构化 |
| 生产监控视频 | 视频监控拍摄的视频 | 生产监控系统 | 非结构化 |


图 2 多源异构数据处理的一般流程

线射频识别数据采集单元模型，可获取零件的状态、时间等实时信息，并在此基础上建立了针对加工工序、工序流、批次与批量的无线射频识别监控模型，从而实现了对离散制造车间生产过程的可视化监控。而对于流程生产制造业中的生产过程数据，主要依靠传感器及上位机对数据进行采集。陈开胜[17]提出了采用分布式控制系统（distributed control system，DCS）和可编程逻辑控制器（programmable logic controller，PLC）等辅助控制系统和控制装置进行数据采集的方法，该方法是对计算机、网络和数据库的综合运用。此外，在流程生产中，以计算机为基础的数据采集系统还有数据采集与监视控制（supervisory control and data acquisition，SCADA）系统。其中，PLC主要应用于生产现场的温度测控；DCS主要应用在对测控精度及速度要求较高的生产现场的数据采集；SCADA则融合了PLC的现场测控功能和DCS的组网通信能力，可以对分散点进行控制，从而实现对分布范围较广的生产现场的覆盖。西门子公司在PLC的基础上加入了网络以及软件等，开发了SIMATIC PCS7西门子SCADA系统、SIMATIC WinCC西门子SCADA系统等控制系统，而DCS厂商霍尼韦尔公司也在其系统中融入了PLC，以增强其逻辑控制，开发了过程知识系统（process knowledge system，PKS）[18]。对于在离散制造业及流程制造业中均广泛存在的日志数据及多媒体数据等，同样根据其各自的特点采用不同的数据采集方法。对于制造生产过程产生的日志数据文件，可以采用Flume这一分布式、高可靠、高可用的日志采集传输系统。陈飞等人[19]提出了一种基于Flume并

结合Elasticsearch及Kibana的新型分布式采集系统，该系统适用于海量日志数据的采集。针对生产过程对音频、视频等多媒体数据的监控，有利用多媒体流处理引擎直接抓取或利用厂商提供的软件开发工具包（software development kit，SDK）开发数据导入程序的数据采集方法[20]。李凤娇[21]在海康威视的8100系列网络硬盘录像机的基础上，通过调用海康威视提供的SDK中的相关接口函数读取实时视频流。另外，浙江宇视科技有限公司的IP流媒体解决方案可以通过安装流媒体服务器软件来对多媒体数据进行实时访问及存储。对于这几种典型的数据采集方法，本文根据其面向的数据类型及在生产过程中的应用进行了总结，见**表2**。

另外，针对数据采集的新需求，相关研究也提出了许多与网络技术相结合的创新型数据采集方法。马吉军等人[22]提出了一种基于边缘计算的生产数据采集方法，利用蜂窝网络对生产设备进行网络化改造，并利用边缘网关对采集到的生产数据进行本地处理。许瀚之和杨小健[23]提出了一种基于虚拟专用网（virtual private network，VPN）的远程工业数据采集系统，在已建好的VPN环境下通过用于过程控制的OLE（OLE for process control，OPC）客户端进行数据采集。

对于目前几种典型的数据采集场景，实际应用中根据其采集的数据类型及要求等，采用Flume、RFID、传感器等不同的采集方法，这些方法具有不同的优势。而面对目前数据量迅速增长以及数据类型日益复杂化的问题，传统数据采集方法难以满足更具实时性、更精确的采集要求，因此，与物联网等前沿技术相结合成为数据采集的发展趋势。

## 3.2 数据集成

多源异构数据集成是整合来自多个数据源的数据，屏蔽数据之间类型和结构上的差异，解决多源异构数据的来源复杂、结构异构问题，从而实现对数据的统一存储、管理和分析，实现用户无差别访问，充分发挥数据的价值。数据集成的关键技术包括数据存储管理、数据清洗与转换及数据降维。

### 3.2.1 数据存储管理

数据的存储管理是多源异构数据处理过程中非常重要的一个环节，选择合理的数据库可以减少数据检索的时间，提高数据查询的准确度，是后续数据处理的基础。目前常见的数据库技术包括：以MySQL、Oracle、DB2、SQL Server等为代表的SQL数据库，以Redis、HBase、

表 2 典型数据采集方法及其在制造业生产过程中的应用

| 比较项 | RFID | 传感器 | Flume | 流媒体服务器 |
|---|---|---|---|---|
| 主要应用场景 | 离散制造业生产过程的设备及产品信息 | 流程制造业生产过程的设备及产品信息 | 生产过程日志数据 | 生产过程监控数据 |
| 数据类型 | 数字量 | 模拟量 | 日志数据文件 | 多媒体数据 |
| 系统交互 | 较多应用在ERP系统中 | 传递给上层的PLC和DCS，最终与上位机对接 | 可以将采集的数据存储到任何集中存储器 | 多媒体流处理引擎实时抓取并上传 |
| 优势 | 读取距离远、准确率高 | 数据类型丰富、应用领域广 | 高可靠性、高可扩展性 | 实时性强、清晰度高 |

MongoDB、Neo4j等为代表的NoSQL数据库，以及NewSQL数据库。

美国甲骨文公司研发的Oracle是一种高效、适应高吞吐量的关系型数据库系统，在数据量大、对系统性能稳定要求高的钢铁[24]、煤炭[25]、汽车制造[26]行业应用广泛。美国IBM公司开发的DB2具有伸缩性能良好、查询性能良好以及向下兼容性好的特点，适用于海量数据的存储管理，在政府、银行等广泛应用，另外在宝钢[27]、本钢等钢铁企业也有应用。制造业生产制造过程中产生的海量多源异构数据包含结构化、半结构化和非结构化多种数据。由于面向结构化数据的传统关系型数据库在伸缩性、容错性、可扩展性等方面存在的固有局限性[28]，单独使用难以满足对海量多源异构数据进行存储管理的要求，因此NoSQL数据库成为目前研究与应用的热点。

根据数据存储模型和特点，NoSQL数据库可分为4种典型类型：以Redis、Memcached为代表的键值存储数据模型，以Bigtable、HBase为代表的列式存储数据模型，以MongoDB为代表的文档存储数据模型，以及以Neo4j为代表的图形存储数据模型。Redis常被应用在社交领域，用来存储用户关系和计数。由于生产过程中多源异构数据对实时性要求较高，因此Redis在制造业数据存储中常被用作缓存系统，以保障数据存储的低时延性。在电力计量采集系统中[29]，基于Redis的分布式写缓存子系统用于缓存采集的计量数据，再批量写入关系数据库。在大型机械设备的数据采集与存储中，熊肖磊等人[30]在数据层基于Redis实现了实时数据的解析缓存，使系统具有高效缓存数据的能力。Google Bigtable开源实现的HBase具有扩展性好、备份机制完善的特征，当制造业生产过程涉及多源异构数据的统计分析时，可使用HBase对来自各个子系统的数据进行同步整合存储。例如，在分布式电源控制系统[31]中，可以实现各个分布式电源系统的运行状态数据至HBase数据库的同步。查询语言功能强大的文档存储数据库MongoDB适合数据量大、数据模型无法确认、需要对接多个数据源等的场景，数据来源复杂是制造业生产过程多源异构数据的主要特点之一，因此MongoDB常被用于多个数据源或子系统的对接。在工业生产中，MongoDB可用于对过程的连续监控[32]；在混凝土行业[33]中，MongoDB用来存储海量的混凝土生产消耗数据，并实现多个系统之间的数据对接；在电力行业[34]，MongoDB可以实现电网图形的多时态、多级分布式存储。

针对工业制造业过程数据产生速率快，实时性要求高，对事务的原子性（atomicity）、一致性（consistency）、隔离性（isolation）、持久性（durability）（即ACID）要求低的特点，冯德伦[35]提出了NoSQL数据库合理组合的工业历史数据存储方案。针对制造业生产过程多源异构数据的来源更加多样化的发展趋势，NoSQL数据库与其他技术相结合的大数据平台或解决方案近年来也有不少案例。赵德基等人[36]提出了基于Dubbo与NoSQL的工业领域大数据平台，针对工业多源异构数据的接收、存储、计算、分析及展示，根据不同场景的业务需求提供了相应的解决方案。文棒棒和曾献辉[37]提出了一种基于传统数据库多表架构与NoSQL大数据库相结合的新型数据存储方案实现实时数据的分布式存储。

除此之外，451 Group的分析师Aslett M[38]提出了NewSQL技术，其具有NoSQL对海量数据的存储管理能力，同时还保持了传统数据库支持ACID和SQL的特性，但目前应用范围大多为专有软件或特定场景。对

于上述几种典型的数据库技术，笔者对数据库模型、支持的数据类型和应用场景等进行了对比，结果见**表3**[39]。

以上几种典型的数据库技术均有其特定的优势及应用场景，而在特定复杂的应用场景中，单一的数据库往往难以满足人们对数据存储管理等多方面的要求，李东奎和鄂海红[40]提出了关系型数据库不能完全被NoSQL数据库替代的观点，并基于Hibernate OGM建立了统一的SQL和NoSQL数据库访问模型，使得两类数据库能够在同一个框架下按照统一的规则进行读写。因此，根据具体的应用场景，选择不同类型的数据库进行混合部署，使数据库之间形成互补，是目前多源异构数据存储管理的发展趋势。

### 3.2.2 数据清洗与转换

准确可靠的数据是进行有效数据分析、数据挖掘的前提。在实际的生产过程中，由于多源异构数据来源众多的特征，采集到的数据的质量难以保证，缺失的、错误的、不一致的等不符合规范的“脏数据”普遍存在，同时来自不同系统的数据的格式也并不统一，这些都会给数据的有效分析带来困难[41]。数据清洗的目的就是检测数据中存在的“脏数据”，通过数据筛选、数据修复等手段提高数据的质量。而数据转换主要是将多源异构数据转换成统一的目标数据格式，并完成对不同数据指标进行转换的计算。

针对生产过程中不同的问题数据，可以给出不同的数据清洗方法。由于制造业生产过程中的多源异构数据往往来自多个数据源，各数据源通常具有不同的数据库系统、接口服务等，因此数据具有结构类型多样、表达形式不统一等特点，这就导致采集的数据中会存在数据缺失、数据错误、数据不一致等问题[42]。对于缺失的数据，大多数情况下需要手工进行填入，某些情况下可以通过统计学习的方法对缺失值进行处理。曹林[43]针对具有聚类特征的数据集，提出了一种回归插补的缺失值清洗框架。对于错误数据，首先利用统计分析的方法对可能出现的错误值进行识别，然后才能对错误数据进行清除，达到数据清洗的目的。对于不一致的数据，可以基于关联数据之间的一致性来检测数据潜在的错误，并进行修复，以完成对多数据源数

**表 3 典型数据库系统 [39] 及其在制造业生产过程中的应用**

| 比较项 | SQL数据库 | NoSQL数据库 | | | | NewSQL数据库 |
|---|---|---|---|---|---|---|
| 数据库模型 | 关系数据库 | 键-值存储 | 列式存储 | 文档存储 | 图形存储 | 关系数据库 |
| 数据库代表名称 | MySQL | Redis | HBase | MongoDB | Neo4j | PostgreSQL |
| 实现语言 | C和C++ | C | Java | C++ | Java | C |
| 是否结构化数据 | 是 | 自由 | 自由 | 自由 | 自由 | 是 |
| 协议类型 | TCP/IP | 基于TCP的文本协议 | RPC协议 | BSON协议 | JSON/REST协议 | TCP/IP |
| 是否支持事务 | ACID | 半支持，乐观锁控制事务 | 支持行级事务 | 不支持 | ACID | ACID |
| 应用场景 | 适用于传统制造业生产过程中的结构化数据 | 适合作为数据缓存系统，以保障生产数据存储的低时延性 | 适合生产过程中有数据设计统计需求的场景 | 适合需要对接多个数据源等场景 | 适用于图形类数据，例如社交网络推荐系统等，在制造业中应用较少 | 适用于某些专有软件及特定场景中的海量数据管理 |

据的清理[44]。

对于制造业生产过程中的多源异构数据来说，单一的数据清洗方法难以满足实际需求，这就需要一个系统的数据清洗方案。ETL（extract、transform、load）工具是一类常用的大数据预处理工具，应用广泛的有国外开源的Kettle工具、IBM公司的Datastage以及Informatica，其在数据清洗环节发挥着十分重要的作用。也有许多研究人员按照不同的需求对ETL技术进行了改进与完善。周瀚章等人[45]设计了一种基于区域划分算法的ETL高效数据清洗方案，解决应用ETL时产生的大量错误属性数据的问题。ETL工具不仅在数据清洗方面具有广泛的应用，同时也是数据转换的主要工具。孙安健等人[46]设计了一种可以屏蔽异构数据源访问差异的通用ETL工具，提供了大量转换组件来灵活处理复杂的应用场景。陈玉东和姚青[47]提出了一种应用于业务流程数据的转换规则，通过设计流程数据转换算法来将流程日志中的数据快速准确地转换成评估系统需要的标准数据。

除此之外，针对不同的制造业门类及数据采集方法，有不同的数据清洗方案。针对RFID采集数据实时性强、数据量大的特点，余杰和王睿[48]提出了基于时间和基于时间间隔的布鲁姆滤波模型，可以在低内存的情况下保证数据应用的实时性。针对生产车间制造物联环境下采集到的数据连续性、冗余性强的特点，蓝波等人[49]提出了一种基于卡尔曼滤波模型的滑动窗口技术，该技术更加适用于RFID标签移动的生产场景。这些研究针对不同的生产制造场景、不同的采集数据类型和特点，对数据清洗方法进行了改进和完善，使其更加适应实际应用的需要。

目前，深度学习和众包技术开始在数据清洗环节得到应用。郝爽等人[50]提出了利用深度学习模型解决复杂数据清洗任务的方法。针对参与者水平参差不齐造成数据清洗质量较低的情况，万耀璘等人[51]提出了在决策阶段利用成熟计算机算法来提高众包可靠性的方案。深度学习可以减轻用户制定数据清洗规则的负担，众包技术将数据清洗任务发送到互联网，利用公众的参与来提高数据清洗的效率，二者与传统数据清洗技术的结合是数据清洗技术在未来一段时间的发展趋势。对于数据转换来说，ETL工具仍然是提高数据质量、屏蔽数据差异的首选工具。因此，对ETL工具自身现有的扩展性差、调试不便利等局限性进行改进和完善是下一步研究与开发的重点。

### 3.2.3 数据降维

多源异构数据具有种类繁多、结构复杂的特点，为了从原始数据中提取更加可靠、有效的数据信息，需要消除无关、冗余的特征，生成新的特征数据，从而实现对高维数据的降维。在现代制造技术的发展中，制造业生产过程中海量的多源异构数据往往维数较高且大量数据之间存在较高的相关性，这给数据降维带来了更高的难度。一般来说，可以通过对数据进行特征选择或者特征提取来实现数据降维。特征选择的方法通过对原始特征集合中的元素进行选择来得到原始特征集合的子集，从而实现降维；而特征提取的方法则通过对不同特征进行组合来得到新的特征集合，从而达到数据降维的目的。

特征选择不改变特征的含义，从原始特征数据集中选择具有代表性和统计意义的特征，以实现降维的目的。特征选择方法包括基于全局搜索、随机搜索以及启发式搜索策略的特征选择方式和基于Filter、Wrapper的特征选择算法。

全局搜索策略遍历原始特征集，通过

评价准则选择满足特定条件的特征子集，其优点是可以得到最优特征子集。但制造业生产过程中的多源异构数据往往是具有多个独立或相关属性的高维数据，因此运算成本较高，在实际中难以应用。随机搜索策略首先随机选择特征，然后用模拟退火算法进行顺序搜索，或用遗传算法进行无规则搜索，再根据分类的有效性对特征赋予权重，选择权重大于定义阈值的特征。由于随机搜索易受随机因素的影响，不确定性较高，不同的参数设置对随机搜索结果也有较大的影响。启发式搜索策略又被称为序贯优选法，可以实现最优特征子集与计算复杂度之间的平衡。相比于前两种方法，其复杂度较低、效率更高。陈建华[52]针对设备故障中对数据集降维的问题，提出了一种基于关联关系与启发式搜索组合的特征选择方法，特征子集通过双向搜索算法产生，并通过计算属性之间的关联关系来剔除冗余属性，提高了效率和准确性。

基于Filter的特征选择直接根据评价准则对数据的统计特征进行评价，去除重要程度低的特征，选出的特征子集一般规模较大，适合作为特征预筛选器。基于Wrapper的特征选择依赖后续分类算法，将子集的选择看作搜索寻优问题，根据分类器的准确率来对特征子集进行评价，其分类效率与精度都较高。制造过程中的多源异构数据往往特征众多且关系复杂，田文荫[53]提出了针对高维制造过程的结合偏最小二乘回归与Wrapper特征选择的混合特征选择方法，同时针对制造业生产数据常出现的类别间不平衡问题，提出了一种基于G-Mean的新的混合特征选择方法，在降维能力和分类性能方面均取得了良好的结果。

特征提取通过将原始特征变换成具有具体物理意义或统计意义的特征，将高维的特征向量变换为低维的特征向量。由于制造业生产过程中的多源异构数据来源于制造生产各个环节中的设备、产品信息等，具有较强的专业性及关联性，因此在进行数据特征提取时会更加注重特征背后的物理意义以及特征之间的关联性。传统的特征提取方法包括线性主成分分析（principal component analysis，PCA）、线性判别分析（linear discriminant analysis，LDA）、独立成分分析（independent component analysis，ICA）、非线性的核主成分分析（kernel principal component analysis，KPCA）、核独立成分分析法（kernel independent component analysis，KICA）。

主成分分析法主要通过观测变量内部的相互关系来整理信息，将可能相关的原始数据集转换成线性不相关的新特征集合，实现高维数据向低维数据的压缩。在纺织业中，刘海军等人[54]利用本色布纹理的自相关性特征，采用主成分分析法去除其相关性，得到了纹理的主成分，将在主成分方向上样本图像的压缩结果作为特征变量，进行分类检测，得到了较高的分类准确度。在煤矿井下供电系统故障检测中，郭凤仪等人[55]通过对时频域变换的回路电流特征矩阵的奇异值进行主成分分析，得到了故障识别的特征，进一步采用遗传算法优化的支持向量机对故障电弧特征的有效性进行测试，可以有效识别电机及变频器负载回路的串联故障电弧。针对机械装备制造业生产过程对加工设备依赖程度高的问题，姚菲[56]提出了一种对备件预测理论的创新性探索，利用基于主成分分析和支持向量机的综合算法进行需求预测，从而实现对设备备件需求的预测。主成分分析法适合处理呈高斯分布的原始数据，但实际生产过程中多源异构数据分布的复杂程度远超高斯分布，这限制了主

成分分析法的应用。

线性判别分析法是有监督的特征提取方法，降维后在新的子空间中使同类特征尽可能接近、不同类特征尽可能分散，与主成分分析法一样，也适合用于处理高斯分布数据。针对模拟电路故障诊断中故障数据的特征提取方法，肖迎群等人[57]对模拟故障数据在主元变换空间进行线性判别分析，并将最优判别特征模式应用于模式分类器，在充分简化模式分类器模型及降低系统运行成本的基础上获得了较好的诊断结果。另外，在图像识别数据分析中，线性判别分析法也是一个十分具有优势的工具。在对铅酸蓄电池X射线图像的特征提取中，杨金堂等人[58]分别采用主成分分析法、线性判别分析法以及二次线性判别分析法，最终得出二次线性判别分析法在该图像识别中具有较高识别率的结论。

独立成分分析法将原始数据分解为若干独立分量的线性组合，更适合用于处理非高斯分布的情况。杨冲等人[59]采用独立成分分析和主成分分析两种常用方法对制浆造纸废水处理过程中的传感器故障进行检测，由于制浆造纸废水处理过程中的数据呈非高斯分布，ICA的整体故障检测率高于PCA。针对滚动轴承在噪声背景下产生故障时的振动信号，姜怀斌[60]利用独立成分分析在数据独立性分析方面的优势，提出了一种独立元核FDA（ICA-KFDA）故障检测模型，提高了故障诊断的准确率，降低了漏检率。

对于图像视频等呈非线性分布的数据，需要使用非线性的特征提取方法。核主成分分析由Scholkopf B等人[61]在PCA的基础上提出，将原始数据通过核函数映射到高维度空间后，再利用PCA进行降维。针对旋转机械结构中轴承状态的识别，谢锋云等人[62]提出了粒子群优化核主成分分析法，对轴承的复合特征集进行特征提取，继而由支持向量机对识别特征集进行识别分类，提高了轴承状态识别的准确率。对于行星齿轮传动系统故障，贺妍和王宗彦[63]用粒子群优化方法改善了核主成分分析法对非线性问题的分析，新方法在行星齿轮磨损程度的识别和诊断中取得了良好的结果。

核独立成分分析法也是利用相同的思想在ICA的基础上进行扩展的，近年来被广泛应用在非线性混叠的源分离技术中。针对旋转机械结构中的滚动轴承故障，刘嘉辉等人[64]提出了一种全矢谱和独立分量分析（ITD和KICA）相结合的盲源分离法，对采样的滚动轴承故障信号进行有效的信噪分离，在降噪的同时能够更加全面、准确地提取信息，并进行轴承故障诊断。针对化工行业的润滑油生产过程，许亮等人[65]提出了基于混合核函数的KICA-LSSVM故障分类方法，提高了故障诊断的速度和准确性。

除了对这些传统的特征提取方法进行优化以外，针对制造业生产过程中数据的特点，一些研究提出了不同的方法对数据特征进行提取。针对生产现场传感器时钟差别及生产设备运行原理导致的不同数据源之间可能存在延迟关联的问题，张守利等人[66]提出了一种面向时延的传感器数据特征提取方法，利用基于皮尔逊相关系数的曲线排齐算法调整不同传感器数据之间的时间，使得调整之后的数据相关性达到最大。苗爱民等人[67]提出了一种基于局部线性嵌入（locally linear embedding，LLE）的非线性故障检测新技术，可以有效地计算出保留了局部邻域结构信息的数据的低维嵌入。尚超等人[68]针对制造生产过程中某些产品质量和关键变量始终难以在线测量的问题，构建了一种基于历史测量数据驱动的软传感器，从而对这些变量进行稳定可靠的在线估计。

随着制造业多源异构数据中非结构化数据所占份额的增多，对多源异构数据的特征提取在数据处理中的重要性也大大增加，而在未来一段时间内，对于多源异构数据处理平台来说，对实时数据以及高维度数据集的特征提取仍然是一个挑战。同时，由于工业生产环境的复杂性，针对工业生产过程中的数据降维，要更多地结合业务场景本身，利用先验知识或者专家知识对数据进行降维。

### 3.3 数据分析

数据分析是多源异构数据处理的关键，是指在数据采集与数据集成环节的基础上对工业生产数据的信息和知识进行提取，其目的是利用数据挖掘、机器学习、统计分析等技术对集成的多源异构数据进行分析和处理，从而提取出有价值的信息和知识，用于检测制造生产运行状况和生产产品质量检测、指导人员做决策等。针对工业生产中的数据分析技术等问题，其他学者也有相关研究[69-70]，但本文从更广的应用领域及更全面的方法的角度对制造业生产过程中的数据处理方法进行综合研究。目前，数据分析环节的关键技术包括关联分析、分类分析和聚类分析等。

#### 3.3.1 关联分析

数据关联分析就是发现表面看来无规律的数据间的关联性，从而发现事物之间的规律性和发展趋势等。常用的关联规则挖掘算法包括Apriori算法和FP-Growth算法。

Apriori算法首先通过遍历数据库确定频繁项集，然后根据支持度阈值进行修剪，最后根据支持度来计算可信度，从而确定关联规则，是一种被广泛应用的关联规则挖掘算法。针对大型化和复杂化的机械装备制造业生产过程中异常事件发生概率高、报警数量巨大的问题，樊虹[71]提出了基于数据挖掘Apriori算法的工业过程报警处理方法，缩小了重复报警的数量，提升了对报警事件的处理效率。但是该算法仍然存在需要频繁遍历数据库从而产生大量候选集的问题。针对这一问题，周凯等人[72]提出了一种仅需对数据库扫描一次即可实现改进Apriori算法，可以有效地提高产生有效频繁项集的效率。除此之外，刘芳和吴广潮[73]提出了一种将数据库转换为矩阵形式，通过缩小候选项集规模、减少无用候选项集生成来提高算法效率的方法。

FP-Growth算法是对Apriori算法最经典的改进，采用频繁模式树（FP-tree）存储频繁项集，减少数据库扫描次数。针对制造业设备对快速准确诊断设备故障的需求，张斌等人[74]提出了一种基于兴趣属性列的改进FP-Growth算法的数据挖掘方法，从而实现对工业生产设备故障的快速准确诊断。针对轮胎制造过程中质量异常的问题，李敏波等人[75]提出了一种改进后的FP-Growth并行算法，该算法能够高效地找到影响轮胎质量的因素。另外，针对FP-Growth算法中存在的FP-tree占据空间过大的问题，顾军华等人[76]通过对FP-Tree的规模大小和计算量以及F-List分组策略进行优化，提出了一种新的基于Spark的并行FP-Growth算法——BFPG算法。

除上述两种数据关联分析算法外，由于制造生产过程中数据量在不断增加，在线的动态数据关联分析具有更加现实的意义。Hidber C[77]提出了一种在线的关联分析数据挖掘算法——CARMA算法，该算法具有在线实现数据关联分析、精度高、允许用户在线调整阈值的优点。此后，于丽等人[78]分别对算法的参数估计、数据集遍

历次数进行了优化改进，提高了算法的速度及精度。如今，CARMA算法在预测和控制领域得到了广泛应用。

目前关联分析方法存在诸多不足，如何利用关联规则算法对非结构化数据进行有效处理、如何将关联规则算法与其他的决策方法结合以实现更准确的数据分析等，均有待进一步的研究和发展。

### 3.3.2 分类分析

对于制造业生产过程的数据分析来说，数据的分类技术是实现数据信息挖掘及结果预测的十分重要的方法之一。

分类是指通过算法将数据划分到已经定义好的类别中。常用的分类算法包括决策树算法、基于规则的分类法、人工神经网络算法、深度学习算法、支持向量机（SVM）算法、贝叶斯算法等。

决策树通过对数据集的分析归纳进行学习，应用范围广泛，对于key-value类型的数据来说是最优选择。目前，较为常见的决策树分类算法有C4.5、SLIQ和SPRINT。决策树算法在生产计划安排方面的应用备受关注。针对离散工业的静态Job Shop调度问题，王成龙[79]提出了用决策树模型提取调度知识的方法，对生产调度方案进行了优化。针对机械装备制造业生产计划中工单加工顺序和同一机器不同工件加工顺序等历史数据，于艺浩[80]提出了一种可根据实时数据为工件安排合适的机器的决策树模型，达到了制造车间根据生产状态实时优化调度的效果。另外，在产品质量检测与分析方面，决策树算法也有非常广泛的应用。针对我国冷轧酸洗产品生产技术尚不成熟、产品表面不合格率较高的问题，郭龙波[81]通过对冷轧酸洗产品数据使用二分决策树等工具进行分析，得出了影响冷轧酸洗产品表面质量缺陷的因素以及判定标准，使企业能够更高效、准确地对产品缺陷进行检测。宋建聪[82]提出了一种基于C4.5决策树算法的生产过程质量分析模型，通过找出引起质量问题的主要因素来对产品质量缺陷进行责任分析和诊断，进而采取针对性的措施来提高产品合格率。

基于规则的分类法是利用用户为每个类直接确定的分类规则来形成类别模板，规则分类器通过统计样本中满足分类规则的规则数和次数来确定样本种类的分类方法，常用来产生更易于解释的描述性模型，更适用于处理类分布不平衡的数据集。在能耗分析系统中，许明洋[83]对基于规则的节能措施实施分类算法的应用进行了分析，基于规则的分类法需要用户自己学习规则，与其他分类算法相比，灵活性与准确性较差。

人工神经网络（artificial neural network，ANN）具有自主学习、容错性高的特点，适合处理模糊、非线性的数据，其中前馈式神经网络模型常用于分类算法。其中，反向传播（back propagation，BP）神经网络算法主要利用反向传播算法对网络的权值和偏差进行反复调整训练，使输出的向量尽可能接近期望向量。但由于其随机获取网络初始权重和阈值的特点，BP神经网络具有收敛时间长、易陷入局部最优解的缺点。周福来[84]、张细政等人[85]、关子奇等人[86]、夏颖怡[87]均基于遗传算法对BP神经网络进行了优化，从而实现了对齿轮设备故障、焊接熔池照度以及刀具寿命等的精确诊断。李世科[88]采用列文伯格-马夸尔特（Levenberg-Marquardt，LM）算法对BP神经网络进行改进，对液压支架顶梁疲劳寿命进行了精确的预测。罗校清[89]应用主元分析法对BP神经网络进行了优化，最终实现了对机械设备故障的准确判断和及时报警。

深度学习最早起源于对人工神经网络的研究，最早由多伦多大学的Hinton G E等人[90]在2006年提出，指基于样本数据的包含多层次的深度网络结构的机器学习过程。深度学习本质上属于机器学习的范畴，是机器学习领域一个新的研究方向，在图像、语音、文本分类识别方面具有非常好的优势[91-92]，具有强大的对不同类型数据的处理能力，因此对制造业生产过程中的数据分析起到非常大的作用。如今被广泛熟知的深度学习基本模型包括深度神经网络（deep neural network，DNN）、循环神经网络（recurrent neural network，RNN）、卷积神经网络（convolutional neural network，CNN）、深度置信网络（deep belief network，DBN）等。深度神经网络可以简单地理解为含有多个隐藏层的神经网络，其优势体现在对无标签数据的自我学习。对于机械设备中常见的传动零件齿轮的故障监测，李嘉琳等人[93]应用深度神经网络来诊断早期齿轮点蚀故障，将采集的振动信号直接作为DNN输入，可以有效解决特征提取环节造成的较大误差，与传统ANN诊断结果相比，故障诊断率得到了提高。针对制造车间中关键刀具设备的寿命预测问题，刘胜辉等人[94]将小波包分析方法得到的结果作为输入来训练深度神经网络，建立刀具剩余寿命预测模型，可对切削刀具剩余寿命进行精确的预测。卷积神经网络是一种包含卷积计算的前馈神经网络，长期以来是图像识别领域的核心算法之一。曹大理等人[95]采用卷积神经网络自适应地提取特征，避免了人为提取的局限性，提高了刀具磨损在线监测的精度。吴志洋等人[96]针对布匹生产中的布匹瑕疵检测，提出了一种基于深度卷积神经网络的单色布匹瑕疵检测算法，很好地解决了人工检测效率低、误检率高的问题。彭大芹等人[97]提出了一种基于卷积神经网络的液晶面板缺陷检测算法，并在传统单向特征融合的基础上提出了双向特征融合的网络结构，提高了检测精度。李广等人[98]针对工业中常见的机床刀具消耗冗余问题，采用异常检测卷积神经网络（CNN-AD）对机床刀具的崩刃进行准确预测。循环神经网络是一类用于处理和预测序列数据的神经网络模型，与传统机器学习方法相比，其对于输入/输出数据没有过多限制，可以用来处理文本、音频和视频等序列数据。针对燃煤电站$NO_x$排放预测模型建模中输入变量特征集确定困难的问题，王文广和赵文杰[99]提出了一种基于数据驱动的门控循环单元（gated recurrent unit，GRU）循环神经网络模型，将GRU作为RNN的神经网络单元，从而使RNN能够分析长时间的时间序列问题，对燃煤电站锅炉$NO_x$排放实现准确预测。对于基于循环神经网络的电力变压器故障诊断模型存在的诊断不清晰、收敛速度慢的缺陷，李俊峰[100]基于蝙蝠算法对循环神经网络的参数进行了优化，改进后的变压器故障诊断模型的收敛性及诊断准确率均得到了较大提升。深度置信网络通过模拟人类大脑对外部信号的处理来实现功能，是由多个限制玻尔兹曼机（restricted Boltzmann machine，RBM）叠加组成的网络模型。王宪保等人[101]运用深度置信网络训练网络的初值，再通过对比重构图像与缺陷图像，实现快速准确的太阳能电池片表面缺陷检测。李梦诗等人[102]提出了一种基于深度置信网络的新型风力发电机故障诊断方法，并通过与传统检测方法进行对比，验证了该算法的鲁棒性。刘浩等人[103]提出了一种基于多参数优化深度置信网络的滚动轴承外圈损伤程度识别方法，可有效地提高故障识别的准确性和稳定性。目前深度学习模型在制造生产数据分析中的大致发展方向是与其他算法相结合，对深度学习基本

模型中的参数、结构进行优化，从而提高算法的精确性与鲁棒性，实现更精准的检测与预测。

支持向量机[104]是一种通过核函数免去高维变换，直接将低维参数代入核函数从而得出高维向量内积的分类方法，常用于故障诊断。针对机械制造业中滚动轴承造成的故障识别问题，吕震宇[105]提出了一种使用磷虾群算法优化的支持向量机，对轴承状态进行精确诊断，从而精确地识别滚动轴承的故障类型，较传统支持向量机的识别精度更高。吕维宗等人[106]提出了基于量子粒子群优化（quantum particle swarm optimization，QPSO）算法优化的相关向量机（relevance vector machine，RVM），并进行故障诊断，相较于支持向量机而言，其更适用于小样本处理和在线故障诊断。

贝叶斯分类算法是在贝叶斯公式的基础上，利用概率统计进行分类计算的方法。其中，朴素贝叶斯分类应用最广泛。制造生产过程中少不了电池寿命与电力故障的问题，Ng S S Y等人[107]针对不同工作环境温度及放电电流情况，提出了用于不同工作状况下电池估计和剩余使用寿命预测的朴素贝叶斯模型。李梦婷等人[108]基于增量式贝叶斯算法，提出了一种实时性在线电路故障诊断方法，可以同时实现在线电路故障诊断的高精确性与高实时性。

目前分类分析方法在工业生产中已经有广泛的应用，尤其是基于机器学习的分类方法。但是现阶段单一的数据分类方法并不具有较高的准确性及可靠性，需要不同算法的融合才能产生较为可靠的数据分类及预测结果。然而不同算法的融合势必会造成系统时延，如何平衡系统的可靠性和实时性是研究的方向之一。另外由于工业生产的特殊性和复杂性，针对同一类分类问题，并没有通用的分类方法可以使用，要得到可靠的分类结果，需要与实际场景、实际业务相结合。同时，如果要得到较为准确的分类结果，分类算法模型的训练数据集需要结合生产领域的经验知识进行相应的特征工程处理。

### 3.3.3 聚类分析

聚类就是将相似的数据归为一类，原则是使每一类数据的相似性最大。常用的聚类算法包括基于划分的聚类方法、基于层次的聚类方法、基于密度的聚类方法和基于模型的聚类方法四大类。

其中，最常用的是*K*-means算法。*K*-means算法是一种基于划分的聚类方法，通过随机选择*K*个数据点作为初始聚类中心，根据特定的距离算法将待聚类的数据集分成*K*簇。娄小芳[109]通过对大量铝工业生产历史能耗数据进行处理分析，运用*K*-means算法等方法分析其规律，以此指导生产部门改进参数，降低能耗。针对酿酒不良发酵行为早期迹象的识别，Urtubia A等人[110]通过对产品中29种成分检测的数据采用*K*-means算法进行聚类分析，获得了不良发酵行为模型，从而实现了对产品质量的认定，减少了早期行为造成的损失。但该算法存在聚类结果受选择的初始聚类中心影响较大、处理大数据时间效率低等缺点。徐健锐和詹永照[111]将改进的*K*-means算法和分布式计算框架Spark结合，提出了大数据下的快速聚类算法SparkKM，该算法既弥补了经典*K*-means算法的不足，又发挥了Spark分布式计算处理速度快的优势。

除此之外，常用的聚类方法还有基于密度的DBSCAN算法、基于层次的BIRCH算法以及基于模型的高斯混合模型（GMM）等。基于密度的DBSCAN算法通过对核心点、边界点和噪声点的标记，将具有密度的区域划分成簇。针对风力发电设

备中故障率最高的齿轮箱和主轴的故障识别问题，林涛等人[112]利用DBSCAN聚类算法对运行数据进行密度聚类，对齿轮箱和主轴的故障进行较准确的诊断。针对电力系统信息安全问题，谢静瑶等人[113]采用启发式的自适应算法对DBSCAN算法的部分参数进行估计，改进了聚类效果，从而提高了信息安全预警分析的准确性。基于层次的BIRCH算法利用树结构进行聚类，适用于数据量大、类别数多的数据处理。对于木材加工中木材缺陷的识别问题，吴东洋和业宁[114]采用BIRCH算法对数据集进行一次扫描即可得到较高的聚类质量，提高了识别准确率。针对食品卫生的HACCP（hazard analysis critical control point）自动分类，叶飞跃等人[115]提出了一种多阈值、多代表点的BIRCH算法，该算法可以适应HACCP分类中各种形状的数据集。基于模型的高斯混合模型是一种融合了参数模型和非参数模型的优势的聚类方法，常被应用在语音识别、图像识别等领域。针对机械结构中易损坏的滚动轴承，龙铭等人[116]提出了一种基于自回归高斯混合模型（AR-GMM）的滚动轴承故障程度评估方法。它以早期无故障轴承振动信号的AR模型特征为基准特征，引入后期轴承振动信号的AR特征，可以监测滚动轴承各种形式的早期故障。针对应用广泛的螺栓连接，王刚等人[117]利用监测区域内螺栓连接结构的各种松动工况的实时数据建立高斯混合模型，基于高斯混合模型的概率密度分布之间的相似度最大准则，可有效判断监测区域螺栓的松紧状态。针对印花织物的表面疵点检测，李敏等人[118]在传统高斯混合背景模型的基础上引入了自适应分块建模的思想，在提高印花织物疵点检测准确率的同时，能有效地处理检测过程中的光照不均和噪声等问题。

数据量的迅速增加使得对大规模数据的分类、聚类成为具有挑战性的研究问题。对于分类算法来说，不同的算法均有其独特的优势以及特定的应用领域。对于聚类算法来说，传统聚类算法经过抽样或降维会损失精确性，而并行聚类算法尽管具有对大数据高效、良好的扩展性等优点，但算法实现较复杂。简单高效、扩展性高的面向大数据且不消耗更多软硬件资源的分类聚类算法是未来的主要研究和优化方向。

## 4 结束语

本文对制造业生产过程中多源异构数据的概念和类型、数据处理的方法和技术进行了较为全面的综述和梳理。将生产过程中的多源异构数据按照数据来源和数据类型进行了分类，对数据处理的整体流程进行了定义，并对数据处理过程中的具体方法、技术及其在生产过程中的具体应用进行了总结分析。

随着工业物联网的快速发展，数据的来源更多，数据结构更加多样化，同时生产过程中信息系统对数据处理的实时性、准确性要求更高，这给多源异构数据的处理带来了巨大的挑战。首先，设备的多样性和复杂性会给数据采集方法、技术带来新的挑战，需要增加更为丰富、可靠、高效的数据采集方法和技术；其次，海量的数据对数据存储技术的容量和效率、精度等提出了更高的要求，也对传统的SQL、NoSQL等数据存储系统的扩展能力提出了更高的要求，综合数据存储系统成为未来发展的趋势；最后，实际生产对数据清洗、降维及数据分析方法和技术的效率和精确度的要求进一步提高。另外，只有性能更高的数据处理分析平台及更高效的数据挖掘算法才能满足大规模多源异构数据

的实时处理与分析要求。另外，随着边缘计算在工业生产过程中的快速应用，面向边缘控制器、边缘网关和边缘云的数据采集、存储、处理和分析的方法和技术的研发将成为重点研究方向。

## 参考文献:


[1] 李少波, 陈永前. 大数据环境下制造业关键技术分析[J]. 电子技术应用, 2017, 43(2): 18-21, 25.
LI S B, CHEN Y Q. Analysis of key technologies of manufacturing industry in big data environment[J]. Electronic Technology Application, 2017, 43(2): 18-21, 25.

[2] McKinsey&Company. Big data: the next frontier for innovation, competition, and productivity[M]. New York: McKinsey Global Institute, 2011: 1-28.

[3] BIRNET E. The making of ENCODE: lessons for big-data projects[J]. Nature, 2012(489): 49-51.

[4] 李黎, 华奎, 姜昀芃, 等. 输电线路多源异构数据处理关键技术研究综述[J]. 广东电力, 2018, 31(8): 124-133.
LI L, HUA K, JIANG Y P, et al. Review of research on key technologies for multi-source heterogeneous data processing of transmission lines[J]. Guangdong Electric Power, 2018, 31(8): 124-133.

[5] 李亢, 李新明, 刘东. 多源异构装备数据集成研究综述[J]. 中国电子科学研究院学报, 2015, 10(2): 162-168.
LI K, LI X M, LIU D. Review of research on multi-source heterogeneous equipment data integration[J]. Journal of China Academy of Electronics and Information Technology, 2015, 10(2): 162-168.

[6] 张春红. 基于XML的异构数据库集成技术研究[J]. 廊坊师范学院学报(自然科学版), 2014, 14(4): 29-30, 43.
ZHANG C H. Research on XML-based heterogeneous database integration technology[J]. Journal of Langfang Teachers College (Natural Science Edition), 2014, 14(4): 29-30, 43.

[7] 陈彦萍, 郭超, 杨为惠. 面向生产过程的异构数据服务描述语言IO-DSDL的设计与实现[J]. 计算机与数字工程, 2018, 46(5): 976-980.
CHEN Y P, GUO C, YANG W H. Design and implementation of production process-oriented heterogeneous data service description language IO-DSDL[J]. Computer and Digital Engineering, 2018, 46(5): 976-980.

[8] 袁爱进, 岳滨楠, 闫鑫, 等. 工业大数据的应用与实践[J]. 大数据, 2017, 3(6): 27-41.
YUAN A J, YUE B N, YAN X, et al. Application and practice of industrial big data[J]. Big Data Research, 2017, 3(6): 27-41.

[9] 顾新建, 代风, 杨青海, 等. 制造业大数据顶层设计的内容和方法(上篇)[J]. 成组技术与生产现代化, 2015, 32(4): 12-17.
GU X J, DAI F, YANG Q H, et al. Contents and methods of top-level design of manufacturing big data (Part 1)[J]. Group Technology and Production Modernization, 2015, 32(4): 12-17.

[10] 徐颖, 李莉. 制造业大数据的发展与展望[J]. 信息与控制, 2018, 47(4): 421-427.
XU Y, LI L. Development and prospect of manufacturing big data[J]. Information and Control, 2018, 47(4): 421-427.

[11] 李涛, 曾春秋, 周武柏, 等. 大数据时代的数据挖掘——从应用的角度看大数据挖掘[J]. 大数据, 2015, 1(4): 57-80.
LI T, ZENG C Q, ZHOU W B, et al. Data mining in the big data era-viewing big data mining from the perspective of applications[J]. Big Data Research, 2015, 1(4): 57-80.

[12] 智能制造时代的工业大数据分析——基于物联网的八大工业大数据与应用场景[J]. 智慧工厂, 2015(11): 42-44.
Industrial big data analysis in the age of intelligent manufacturing: eight industrial big data and application scenarios based

on the internet of things[J]. Smart Factory, 2015(11): 42–44.

[13] 张洋洋, 陈进. 基于RFID的离散制造车间实时数据采集系统的设计与实现[J]. 江南大学学报(自然科学版), 2013, 12(1): 54–58.
ZHANG Y Y, CHEN J. Design and implementation of RFID-based real-time data acquisition system for discrete manufacturing workshop[J]. Journal of Jiangnan University (Natural Science Edition), 2013, 12(1): 54–58.

[14] 商秀芹, 李梦瑶, 熊刚, 等. 面向孵化器行业的云计算与大数据服务平台[J]. 软件, 2017, 38(6): 1–6.
SHANG X Q, LI M Y, XIONG G, et al. Cloud computing and big data service platform for incubator industry[J]. Software, 2017, 38(6): 1–6.

[15] 许周祥, 陈绪兵, 王瑜辉, 等. RFID技术在智能化生产线中的应用[J]. 机械工程与自动化, 2017(4): 138–139, 141.
XU Z X, CHEN X B, WANG Y H, et al. Application of RFID technology in intelligent production line[J]. Mechanical Engineering and Automation, 2017(4): 138–139, 141.

[16] 曹伟, 江平宇, 江开勇, 等. 基于RFID技术的离散制造车间实时数据采集与可视化监控方法[J]. 计算机集成制造系统, 2017, 23(2): 273–284.
CAO W, JIAGN P Y, JIANG K Y, et al. Real-time data acquisition and visual monitoring method for discrete manufacturing workshop based on RFID technology[J]. Computer Integrated Manufacturing System, 2017, 23(2): 273–284.

[17] 陈开胜. 制造业数据采集技术探究[J]. 开封大学学报, 2017, 31(2): 93–96.
CHEN K S. Research on manufacturing industry data acquisition technology[J]. Journal of Kaifeng University, 2017, 31(2): 93–96.

[18] 刘少锋, 陈晓艳, 张宇辉. 霍尼韦尔SCADA系统在城市燃气管网中的应用[J]. 江苏科技信息, 2017(12): 56–57, 60.
LIU S F, CHEN X Y, ZHANG Y H. Application of Honeywell SCADA system in urban gas pipeline network[J]. Jiangsu Science and Technology Information, 2017(12): 56–57, 60.

[19] 陈飞, 艾中良. 基于Flume的分布式日志采集分析系统设计与实现[J]. 软件, 2016, 37(12): 82–88.
CHEN F, AI Z L. Design and implementation of a distributed log collection and analysis system based on flume[J]. Software, 2016, 37(12): 82–88.

[20] 刘岩, 王华, 秦叶阳, 等. 智慧城市多源异构大数据处理框架[J]. 大数据, 2017, 3(1): 51–60.
LIU Y, WANG H, QIN Y Y, et al. Multi-source heterogeneous big data processing framework for smart cities[J]. Big Data Research, 2017, 3(1): 51–60.

[21] 李凤娇. 基于海康视频监控系统的目标检测和跟踪[D]. 济南: 济南大学, 2014.
LI F J. Target detection and tracking based on Haikang video surveillance system[D]. Jinan: Jinan University, 2014.

[22] 马吉军, 贾雪琴, 寿颜波, 等. 基于边缘计算的工业数据采集[J]. 信息技术与网络安全, 2018, 37(4): 91–93.
MA J J, JIA X Q, SHOU Y B, et al. Industrial data acquisition based on edge computing[J]. Information Technology and Network Security, 2018, 37(4): 91–93.

[23] 许瀚之, 杨小健. 基于VPN的远程工业数据采集解决方案的实现与设计[J]. 上海交通大学学报, 2016, 50(12): 1866–1872, 1888.
XU H Z, YANG X J. Implementation and design of a remote industrial data acquisition solution based on VPN[J]. Journal of Shanghai Jiaotong University, 2016, 50(12): 1866–1872, 1888.

[24] 李若新. Oracle数据库技术在钢铁企业中的一般应用[J]. 数字通信世界, 2019(5): 192.
LI R X. General application of oracle database technology in iron and steel enterprises[J]. World of Digital Communications, 2019(5): 192.

[25] 刘建庆. 探讨煤炭企业中ORACLE数据库的应用[J]. 电子世界, 2015(18): 43–44.
LIU J Q. Discussion on the application of

ORACLE database in coal enterprises[J]. Electronic World, 2015(18): 43–44.

[26] 郭宇. 大数据时代下Oracle数据库在汽车制造业MIS系统中的应用[J]. 计算机光盘软件与应用, 2015, 18(1): 169–170.
GUO Y. Application of oracle database in MIS system of automobile manufacturing industry in the big data era[J]. Computer CD–ROM Software and Application, 2015, 18(1): 169–170.

[27] 沈波. DB2数据库在宝钢炼焦自动化的应用和实践[C]//全国冶金自动化信息网2016年会论文集. 出版地未知: 出版者未知, 2016.
SHEN B. Application and practice of DB2 database in Baosteel coking automation[C]//The 2016 Annual Meeting of National Metallurgical Automation Information Network. [S.l.: s.n.], 2016.

[28] 申德荣, 于戈, 王习特, 等. 支持大数据管理的NoSQL系统研究综述[J]. 软件学报, 2013, 24(8): 1786–1803.
SHEN D R, YU G, WANG X T, et al. Review of research on NoSQL system supporting big data management[J]. Journal of Software, 2013, 24(8): 1786–1803.

[29] 陈森利, 吴福疆, 林洪浩, 等. 电力计量采集系统中分布式缓存系统研究[J]. 信息技术, 2014(7): 70–73, 77.
CHEN S L, WU F J, LIN H H, et al. Research on distributed cache system in power measurement acquisition system[J]. Information Technology, 2014(7): 70–73, 77.

[30] 熊肖磊, 王春伟, 赵炯, 等. 基于Redis与SSM的大型设备数据运用系统设计[J]. 现代机械, 2018(6): 29–34.
XIONG X L, WANG C W, ZHAO J, et al. Design of large equipment data application system based on Redis and SSM[J]. Modern Machinery, 2018(6): 29–34.

[31] 孟云侠. 基于HBase的分布式电源控制系统研究[J]. 电源技术, 2017, 41(9): 1366–1368.
MENG Y X. Research on distributed power control system based on HBase[J]. Power Technology, 2017, 41(9): 1366–1368.

[32] 冯德伦. MongoDB在存储与分析工业时间序列数据中的应用[J]. 自动化与仪器仪表, 2018(9): 141–144.
FENG D L. Application of MongoDB in storage and analysis of industrial time series data[J]. Automation and Instrumentation, 2018(9): 141–144.

[33] 任会民, 杨旭辉, 刘宪红, 等. 关于混凝土行业MongoDB数据库应用的研究[J]. 科技与创新, 2018(20): 38–39, 42.
REN H M, YANG X H, LIU X H, et al. Research on application of MongoDB database in concrete industry[J]. Science and Innovation, 2018(20): 38–39, 42.

[34] 赵越, 李培, 王震, 等. 电网图形数据管理MongoDB数据库的应用[J]. 计算机系统应用, 2017, 26(3): 239–243.
ZHAO Y, LI P, WNAG Z, et al. Application of MongoDB database for graphic data management of power grid[J]. Application of Computer Systems, 2017, 26(3): 239–243.

[35] 冯德伦. 一种以NoSQL数据库为核心的工业历史数据存储方案[J]. 自动化与仪器仪表, 2018(8): 60–63.
FENG D L. An industrial historical data storage solution based on NoSQL database[J]. Automation and Instrumentation, 2018(8): 60–63.

[36] 赵德基, 王力, 狄军峰. 基于Dubbo+NoSQL的工业领域大数据平台研究[J]. 数字技术与应用, 2017(7): 64–67.
ZHAO D J, WANG L, DI J F. Research on big data platform in industrial field based on Dubbo+NoSQL[J]. Digital Technology and Application, 2017(7): 64–67.

[37] 文棒棒, 曾献辉. 面向工业4.0的多表架构与NoSQL大数据集成的数据存储策略研究[J]. 微型机与应用, 2016, 35(18): 6–9.
WEN B B, ZENG X H. Research on data storage strategies for multi–table architecture and NoSQL big data integration for Industry 4.0[J]. Microcomputer & Application, 2016, 35(18): 6–9.

[38] ASLETT M. What's really new with NewSQL?[J]. ACM, 2016, 45(2): 45–55.

[39] 雷宇辉, 钟雯, 何清, 等. NoSQL数据库研究文献综述[J]. 电子世界, 2017(4): 11–12.

LEI Y H, ZHONG W, HE Q, et al. Literature review of NoSQL database research[J]. Electronic World, 2017(4): 11–12.

[40] 李东奎, 鄂海红. 基于Hibernate OGM的SQL与NoSQL数据库的统一访问模型的设计与实现[J]. 软件, 2016, 37(11): 14–18.
LI D K, E H H. Design and implementation of unified access model for SQL and NoSQL databases based on Hibernate OGM[J]. Software, 2016, 37(11): 14–18.

[41] 陈彤. 多源异构海量石油数据的数据清洗技术研究[D]. 青岛: 中国石油大学(华东), 2017.
CHEN T. Research on data cleaning technology of multi-source heterogeneous massive oil data[D]. Qingdao: China University of Petroleum (East China), 2017.

[42] 杨尚林. 基于机器学习的多源异构大数据清洗技术研究[D]. 南宁: 广西大学, 2017.
YANG S L. Research on multi-source heterogeneous big data cleaning technology based on machine learning[D]. Nanning: Guangxi University, 2017.

[43] 曹林. 基于统计学习的数据预处理缺失值清洗方法研究[D]. 哈尔滨: 哈尔滨工程大学, 2012.
CAO L. Research on data preprocessing missing value cleaning method based on statistical learning[D]. Harbin: Harbin Engineering University, 2012.

[44] 杜岳峰, 申德荣, 聂铁铮, 等. 基于关联数据的一致性和时效性清洗方法[J]. 计算机学报, 2017, 40(1): 92–106.
DU Y F, SHEN D R, NIE T Z, et al. Consistency and timeliness cleaning method based on connected data[J]. Chinese Journal of Computers, 2017, 40(1): 92–106.

[45] 周瀚章, 冯广, 龚旭辉, 等. 基于大数据的ETL中的数据清洗方案研究[J]. 工业控制计算机, 2018, 31(12): 108–110.
ZHOU H Z, FENG G, GONG X H, et al. Research on data cleaning scheme in ETL based on big data[J]. Industrial Control Computer, 2018, 31(12): 108–110.

[46] 孙安健, 王星, 闫晓瑜. 通用ETL工具的研究与实现[J]. 计算机应用与软件, 2012, 29(12): 175–178, 210.
SUN A J, WANG X, YAN X Y. Research and implementation of general ETL tools[J]. Journal of Computer Applications and Software, 2012, 29(12): 175–178, 210.

[47] 陈玉东, 姚青. 基于商务智能的流程评估系统中ETL的研究[J]. 计算机工程与设计, 2014, 35(8): 2752–2756.
CHEN Y D, YAO Q. Research on ETL in process evaluation system based on business intelligence[J]. Computer Engineering and Design, 2014, 35(8): 2752–2756.

[48] 余杰, 王睿. 面向离散制造的RFID数据清洗方法研究[J]. 制造业自动化, 2018, 40(6): 86–89, 122.
YU J, WANG R. Research on RFID data cleaning method for discrete manufacturing[J]. Manufacturing Automation, 2018, 40(6): 86–89, 122.

[49] 蓝波, 吴昊, 王一泽, 等. 基于制造物联的生产数据采集与应用技术研究[J]. 电子设计工程, 2017, 25(17): 21–25, 30.
LAN B, WU H, WANG Y Z, et al. Research on production data acquisition and application technology based on manufacturing Internet of things[J]. Electronic Design Engineering, 2017, 25(17): 21–25, 30.

[50] 郝爽, 李国良, 冯建华, 等. 结构化数据清洗技术综述[J]. 清华大学学报(自然科学版), 2018, 58(12): 1037–1050.
HAO S, LI G L, FENG J H, et al. Overview of structured data cleaning technology[J]. Journal of Tsinghua University (Science and Technology), 2018, 58(12): 1037–1050.

[51] 万耀璘, 徐晴雯, 廖彬超, 等. 众包在城市规划的应用与展望[J]. 清华大学学报(自然科学版), 2019(5): 1–8.
WAN Y L, XU Q W, LIAO B C, et al. Application and prospect of crowdsourcing in urban planning[J]. Journal of Tsinghua University (Science and Technology), 2019(5): 1–8.

[52] 陈建华. 基于关联关系与启发式搜索的特征选择在银行设备故障定位中的应用[J]. 北京:

中国科技论文在线, 2014.
CHEN J H. Application of feature selection based on association and heuristic search in bank equipment fault location[J]. Beijing: China Science and Technology Papers Online, 2014.

[53] 田文萌. 基于特征选择的产品关键质量特征识别方法研究[D]. 天津: 天津大学, 2013.
TIAN W M. Research on product key quality feature recognition method based on feature selection[D]. Tianjin: Tianjin University, 2013.

[54] 刘海军, 单维锋, 张莉丽, 等. 基于主成分分析法的本色布疵点分类算法[J]. 毛纺科技, 2019, 47(2): 70–73.
LIU H J, SHAN W F, ZHANG L L, et al. Classification algorithm of natural fabric defects based on principal component analysis[J]. Woolen Textile Technology, 2019, 47(2): 70–73.

[55] 郭凤仪, 高洪鑫, 王智勇, 等. 基于ST-SVD-PCA的串联故障电弧特征提取方法[J]. 煤炭学报, 2018, 43(3): 888–896.
GUO F Y, GAO H X, WANG Z Y, et al. Feature extraction method of series fault arc based on ST-SVD-PCA[J]. Journal of China Coal Society, 2018, 43(3): 888–896.

[56] 姚菲. 制造业备件库存管理优化体系研究与应用[D]. 北京: 北京邮电大学, 2014.
YAO F. Research and application of manufacturing spare parts inventory management optimization system[D]. Beijing: Beijing University of Posts and Telecommunications, 2014.

[57] 肖迎群, 何怡刚, 刘继乾, 等. 基于主元和判别集成分析的模拟电路故障诊断[J]. 控制与决策, 2015, 30(7): 1321–1324.
XIAO Y Q, HE Y G, LIU J Q, et al. Fault diagnosis of analog circuits based on principal component and discriminant integration analysis[J]. Control and Decision, 2015, 30(7): 1321–1324.

[58] 杨金堂, 林孝毅, 杨正群, 等. 废旧铅酸蓄电池的X射线图像识别分类研究[J]. 机械设计与制造, 2017(10): 156–158, 163.
YANG J T, LIN X Y, YANG Z Q, et al. The study of X-ray image recognition and classification of used lead-acid batteries[J]. Machinery Design & Manufacture, 2017(10): 156–158, 163.

[59] 杨冲, 宋留, 刘鸿斌. 基于独立元分析的制浆造纸废水处理过程故障检测[J]. 中国造纸学报, 2019, 34(1): 66–72.
YANG C, SONG L, LIU H B. Fault detection of pulp and papermaking wastewater treatment process based on independent element analysis[J]. Journal of China Paper Society, 2019, 34(1): 66–72.

[60] 姜怀斌. 基于Fisher判别分析的间歇过程故障诊断研究[D]. 哈尔滨: 哈尔滨理工大学, 2018.
JIANG H B. Research on fault diagnosis of batch process based on fisher discriminant analysis[D]. Harbin: Harbin University of Science and Technology, 2018.

[61] SCHOLKOPF B, SMOLA A, MULLER K R. Nonlinear component analysis as a kernel eigenvalue problem[J]. Neural Computation, 1998, 10(5): 1299–1319.

[62] 谢锋云, 陈红年, 谢三毛, 等. 基于粒子群优化核主元分析的轴承状态识别[J]. 测控技术, 2018, 37(3): 28–31, 35.
XIE F Y, CHEN H N, XIE S M, et al. Bearing state recognition based on particle swarm optimization kernel principal component analysis[J]. Measurement and Control Technology, 2018, 37(3): 28–31, 35.

[63] 贺妍, 王宗彦. 基于PSO-FC优化KPCA的特征提取及行星齿轮磨损损伤程度识别[J]. 机械传动, 2019, 43(2): 137–143.
HE Y, WANG Z Y. Feature extraction of KPCA based on PSO-FC optimization and recognition of wear and damage of planetary gears[J]. Mechanical Transmission, 2019, 43(2): 137–143.

[64] 刘嘉辉, 董辛旻, 李剑飞. 基于ITD-KICA盲分离降噪的滚动轴承故障特征提取[J]. 机械传动, 2018, 42(1): 83–87.
LIU J H, DONG X M, LI J F. Feature extraction of rolling bearing faults based on ITD-KICA blind separation and noise

reduction[J]. Mechanical Transmission, 2018, 42(1): 83–87.

[65] 许亮, 程良伦, 黄志平. 基于混合函数的KICA–LSSVM故障分类方法及应用[J]. 化工自动化及仪表, 2010, 37(3): 14–18.
XU L, CHENG L L, HUANG Z P. KICA–LSSVM fault classification method based on mixed function and its application[J]. Chemical Industry Automation and Instruments, 2010, 37(3): 14–18.

[66] 张守利, 苏申, 刘晨, 等. 面向发电设备预测性维护的传感数据特征抽取方法[J]. 太原理工大学学报, 2018, 49(1): 79–85.
ZHANG S L, SU S, LIU C, et al. Feature extraction method of sensor data for predictive maintenance of power generation equipment[J]. Journal of Taiyuan University of Technology, 2018, 49(1): 79–85.

[67] MIAO A M, SONG Z H, GE Z Q, et al. Nonlinear fault detection based on locally linear embedding[J]. Journal of Control Theory and Applications, 2013, 11(4): 615–622.

[68] SHANG C, YANG F, HUANG D X, et al. Data–driven soft sensor development based on deep learning technique[J]. Journal of Process Control, 2014, 24(3).

[69] 王宏志, 梁志宇, 李建中, 等. 工业大数据分析综述: 模型与算法[J]. 大数据, 2018, 4(5): 62–79.
WANG H Z, LIANG Z Y, LI J Z, et al. Survey on industrial big data analysis: models and algorithms[J]. Big Data Research, 2018, 4(5): 62–79.

[70] 梁志宇, 王宏志, 李建中, 等. 制造业中的大数据分析技术应用研究综述[J]. 机械, 2018, 45(6): 1–13.
LIANG Z Y, WANG H Z, LI J Z, et al. A review on the application of big data analysis in manufacturing industry[J]. Machinery, 2018, 45(6): 1–13.

[71] 樊虹. 工业过程报警的关联规则挖掘方法及应用[D]. 北京: 北京化工大学, 2016.
FAN H. Mining method and application of association rules for industrial process alarms[D]. Beijing: Beijing University of Chemical Technology, 2016.

[72] 周凯, 顾洪博, 李爱国. 基于关联规则挖掘Apriori算法的改进算法[J]. 陕西理工大学学报(自然科学版), 2018, 34(5): 40–44.
ZHOU K, GU H B, LI A G. Improved Apriori algorithm based on association rule mining[J]. Journal of Shaanxi University of Technology (Natural Science Edition), 2018, 34(5): 40–44.

[73] 刘芳, 吴广潮. 一种基于压缩矩阵的改进Apriori算法[J]. 山东大学学报(工学版), 2018, 48(6): 82–88.
LIU F, WU G C. An improved Apriori algorithm based on compression matrix[J]. Journal of Shandong University (Engineering Science), 2018, 48(6): 82–88.

[74] 张斌, 滕俊杰, 满毅. 改进的并行FP–Growth算法在工业设备故障诊断中的应用研究[J]. 计算机科学, 2018, 45(S1): 508–512.
ZHANG B, TENG J J, MAN Y. Application of improved parallel fp–growth algorithm in fault diagnosis of industrial equipment[J]. Journal of Frontiers of Computer Science, 2018, 45(S1): 508–512.

[75] 李敏波, 丁铎, 易泳. 基于FP–Growth改进算法的轮胎质量数据分析[J]. 中国机械工程, 2019, 30(2): 244–251.
LI M B, DING D, YI Y. Analysis of tire quality data based on FP–Growth improved algorithm[J]. China Mechanical Engineering, 2019, 30(2): 244–251.

[76] 顾军华, 武君艳, 许馨匀, 等. 基于Spark的并行FP–Growth算法优化及实现[J]. 计算机应用, 2018, 38(11): 3069–3074.
GU J H, WU J Y, XU X Y, et al. Optimization and implementation of parallel FP–Growth algorithm based on Spark[J]. Journal of Computer Applications, 2018, 38(11): 3069–3074.

[77] HIDBER C. Online association rule mining[C]//ACM SIGMOD International conference on Management of Data. New York: ACM Press, 1999.

[78] 于丽, 刘艳君, 丁锋. CARMA模型多新息增广随机梯度参数估计算法的收敛性[J]. 系统工程与电子技术, 2009, 31(6): 1446–1449.

YU L, LIU Y J, DING F. Convergence of multi-innovation augmented stochastic gradient parameter estimation algorithm for CARMA model[J]. Systems Engineering and Electronics, 2009, 31(6): 1446-1449.

[79] 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 production scheduling based on data mining technology[D]. Hangzhou: Zhejiang University, 2015.

[80] 于艺浩. 基于数据的车间实时调度系统的研究与开发[D]. 沈阳: 沈阳工业大学, 2013.
YU Y H. Research and development of real-time scheduling system based on data[D]. Shenyang: Shenyang University of Technology, 2013.

[81] 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. Analysis of cold rolled surface quality defects based on data mining method[D]. Maanshan: Anhui University of Technology, 2012.

[82] 宋建聪. 数据挖掘在装备制造业质量管理中的应用研究[D]. 广州: 广东工业大学, 2013.
SONG J C. Research on the application of data mining in equipment manufacturing quality management[D]. Guangzhou: Guangdong University of Technology, 2013.

[83] 许明洋. 分类算法在能耗分析系统中的应用场景研究及实现[D]. 北京: 北京邮电大学, 2016.
XU M Y. Research and implementation of classification algorithm in energy consumption analysis system[D]. Beijing: Beijing University of Posts and Telecommunications, 2016.

[84] 周福来. 基于BP神经网络的齿轮设备故障诊断应用[J]. 电子技术与软件工程, 2019(10): 139-141.
ZHOU F L. Application of gear device fault diagnosis based on BP neural network[J]. Electronic Technology and Software Engineering, 2019(10): 139-141.

[85] 张细政, 郑亮, 刘志华. 基于遗传算法优化BP神经网络的风机齿轮箱故障诊断[J]. 湖南工程学院学报(自然科学版), 2018, 28(3): 1-6.
ZHANG X Z, ZHENG L, LIU Z H. Fault diagnosis of fan gearbox based on genetic algorithm to optimize BP neural network[J]. Journal of Hunan Institute of Engineering (Natural Science Edition), 2018, 28(3): 1-6.

[86] 关子奇, 朱玉龙, 刘晓光, 等. 基于GA优化BP神经网络的焊接熔池照度建模[J]. 热加工工艺, 2019, 48(7): 216-220, 223.
GUAN Z Q, ZHU Y L, LIU X G, et al. Modeling of welding pool illumination based on GA optimized BP neural network[J]. Hot Working Technology, 2019, 48(7): 216-220, 223.

[87] 夏颖怡. 基于GA-BP神经网络的刀具寿命预测研究[J]. 精密制造与自动化, 2017(2): 9-11.
XIA Y Y. Research on tool life prediction based on GA-BP neural network [J]. Precision Manufacturing and Automation, 2017(2): 9-11.

[88] 李世科. 基于LM-BP神经网络的液压支架顶梁疲劳寿命预测及应用[J]. 中国矿业, 2019, 28(5): 92-96.
LI S K. Fatigue life prediction and application of the top support of hydraulic support based on LM-BP neural network[J]. China Mining Industry, 2019, 28(5): 92-96.

[89] 罗校清. 基于人工神经网络的工业机械故障诊断优化方法研究[J]. 科技创新与应用, 2017(30): 106-107, 110.
LUO X Q. Research on optimization method of industrial machinery fault diagnosis based on artificial neural network[J]. Science & Technology Innovation and Application, 2017(30): 106-107, 110.

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

[91] JI S W, XU W, YANG M, et al. 3D convolutional neural networks for human action recognition[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence.

Piscataway: IEEE Press, 2013.

[92] DAHL G E, YU D, DENG L. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition[C]//IEEE Transactions on Audio Speech and Language Processing. Piscataway: IEEE Press, 2012.

[93] 李嘉琳, 何巍华, 曲永志. PSO优化深度神经网络诊断齿轮早期点蚀故障[J]. 东北大学学报(自然科学版), 2019, 40(7): 974-979.
LI J L, HE W H, QU Y Z. PSO-optimized deep neural network for diagnosis of early pitting corrosion of gears[J]. Journal of Northeastern University (Natural Science), 2019, 40(7): 974-979.

[94] 刘胜辉, 张人敬, 张淑丽, 等. 基于深度神经网络的切削刀具剩余寿命预测[J]. 哈尔滨理工大学学报, 2019, 24(3): 1-8.
LIU S H, ZHANG R J, ZHANG S L, et al. Prediction of remaining life of cutting tools based on deep neural networks[J]. Journal of Harbin University of Science and Technology, 2019, 24(3): 1-8.

[95] 曹大理, 孙惠斌, 张纪铎, 等. 基于卷积神经网络的刀具磨损在线监测[J]. 计算机集成制造系统, 2020(1): 1-12.
CAO D L, SUN H B, ZHANG J D, et al. Tool wear online monitoring based on convolutional neural network[J]. Computer Integrated Manufacturing System, 2020(1): 1-12.

[96] 吴志洋, 卓勇, 李军, 等. 基于卷积神经网络的单色布匹瑕疵快速检测算法[J]. 计算机辅助设计与图形学学报, 2018, 30(12): 2262-2270.
WU Z Y, ZHUO Y, LI J, et al. A fast algorithm for detecting defects in monochrome cloths based on convolutional neural networks[J]. Journal of Computer-Aided Design & Computer Graphics, 2018, 30(12): 2262-2270.

[97] 彭大芹, 刘恒, 许国良, 等. 基于双向特征融合卷积神经网络的液晶面板缺陷检测算法[J]. 广东通信技术, 2019, 39(4): 66-73.
PENG D Q, LIU H, XU G L, et al. Defect detection algorithm of liquid crystal panel based on bidirectional feature fusion convolution neural network[J]. Guangdong Communication Technology, 2019, 39(4): 66-73.

[98] 李广, 杨欣. 结合深度学习的工业大数据应用研究[J]. 大数据, 2018, 4(5): 6-17.
LI G, YANG X. An industrial big data application research using deep learning[J]. Big Data Research, 2018, 4(5): 6-17.

[99] 王文广, 赵文杰. 基于GRU神经网络的燃煤电站$NO_x$排放预测模型[J]. 华北电力大学学报(自然科学版), 2020(1): 1-9.
WANG W G, ZHAO W J. Prediction model of $NO_x$ emissions from coal-fired power stations based on GRU neural network[J]. Journal of North China Electric Power University(Natural Science Edition), 2020(1): 1-9.

[100] 李俊峰. 基于循环神经网络和蝙蝠算法的变压器故障诊断[J]. 电工技术, 2018(20): 38-41.
LI J F. Transformer fault diagnosis based on recurrent neural network and bat algorithm[J]. Electrical Engineering Technology, 2018(20): 38-41.

[101] 王宪保, 李洁, 姚明海, 等. 基于深度学习的太阳能电池片表面缺陷检测方法[J]. 模式识别与人工智能, 2014, 27(6): 517-523.
WANG X B, LI J, YAO M H, et al. Surface defect detection method of solar cells based on deep learning[J]. Pattern Recognition and Artificial Intelligence, 2014, 27(6): 517-523.

[102] 李梦诗, 余达, 陈子明, 等. 基于深度置信网络的风力发电机故障诊断方法[J]. 电机与控制学报, 2019, 23(2): 114-122.
LI M S, YU D, CHEN Z M, et al. Fault diagnosis method for wind turbines based on deep confidence network[J]. Journal of Electrical Engineering and Control, 2019, 23(2): 114-122.

[103] 刘浩, 熊炘, 周辰, 等. 多参数优化深度置信网络的滚动轴承外圈损伤程度识别[J]. 轴承, 2018(12): 43-48.
LIU H, XIONG X, ZHOU C, et al. Identification of damage degree of outer ring of rolling bearing based on multi-parameter optimized deep confidence

network[J]. Bearing, 2018(12): 43–48.

[104] KHALILIA M, CHAKRABORTY S, POPESCU M. Predicting disease risks from highly imbalanced data using random forest[J]. BMC Medical Informatics and Decision Making, 2011, 11(1): 51–63.

[105] 吕震宇. 磷虾算法优化多分类支持向量机的轴承故障诊断[J]. 制造技术与机床, 2019(5): 130–136.
LYU Z Y. Multi–class support vector machine optimized by Krilling algorithm for bearing fault diagnosis[J]. Manufacturing Technology and Machine Tool, 2019(5): 130–136.

[106] 吕维宗, 王海瑞, 舒捷. 量子粒子群算法优化相关向量机的轴承故障诊断[J]. 计算机应用与软件, 2019, 36(1): 6–11, 16.
LYU W Z, WANG H R, SHU J. Bearing fault diagnosis of correlation vector machine optimized by quantum particle swarm optimization[J]. Computer Applications and Software, 2019, 36(1): 6–11, 16.

[107] NG S S Y, XING Y J, TSUI K L. A naive Bayes model for robust remaining useful life prediction of lithium–ion battery[J]. Applied Energy, 2014: 118.

[108] 李梦婷, 赵帅, 陈绍炜, 等. 基于增量贝叶斯学习模型的在线电路故障诊断[J]. 计算机应用与软件, 2018, 35(6): 70–75.
LI M T, ZHAO S, CHEN S W, et al. Online circuit fault diagnosis based on incremental Bayesian learning model[J]. Journal of Computer Applications and Software, 2018, 35(6): 70–75.

[109] 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 and consumption reduction of aluminum industry production based on pattern recognition and data mining[D]. Changsha: National University of Defense Technology, 2010.

[110] 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): 1510–1517.

[111] 徐健锐, 詹永照. 基于Spark的改进*K*–means快速聚类算法[J]. 江苏大学学报(自然科学版), 2018, 39(3): 316–323.
XU J R, ZHAN Y Z. Improved *K*–means fast clustering algorithm based on Spark[J]. Journal of Jiangsu University (Natural Science Edition), 2018, 39(3): 316–323.

[112] 林涛, 马同宽, 秦冬阳, 等. 基于改进DBSCAN算法的风机故障诊断研究[J]. 现代电子技术, 2018, 41(21): 146–149, 155.
LIN T, MA T K, QIN D Y, et al. Research on fan fault diagnosis based on improved DBSCAN algorithm[J]. Modern Electronics Technology, 2018, 41(21): 146–149, 155.

[113] 谢静瑶, 解思江, 焦阳, 等. 一种改进的启发式自适应DBSCAN聚类算法的研究及其在电力系统信息安全预警分析中的应用[J]. 电信科学, 2017, 33(S1): 117–122.
XIE J Y, XIE S J, JIAO Y, et al. Research on an improved heuristic adaptive DBSCAN clustering algorithm and its application in early warning analysis of power system information security[J]. Telecommunications Science, 2017, 33(S1): 117–122.

[114] 吴东洋, 业宁. 基于BIRCH的木材缺陷识别[J]. 山东大学学报(工学版), 2010, 40(5): 137–140.
WU D Y, YE N. Wood defect recognition based on BIRCH[J]. Journal of Shandong University (Engineering Science Edition), 2010, 40(5): 137–140.

[115] 绍彬, 叶飞跃, 刘佰强, 等. 食品HACCP分类的BIRCH算法[J]. 计算机工程, 2008, 34(23): 59–61.
SHAO B, YE F Y, LIU B Q, et al. BIRCH algorithm for food HACCP classification[J]. Computer Engineering, 2008, 34(23): 59–61.

[116] 龙铭, 文章, 黄文艺, 等. 滚动轴承故障程度评估的AR–GMM方法[J]. 机械科学与技术, 2016, 35(8): 1183–1188.
LONG M, WEN Z, HUANG W Y, et al.

AR-GMM method for evaluating the degree of failure of rolling bearings[J]. Mechanical Science and Technology, 2016, 35(8): 1183-1188.

[117] 王刚, 肖黎, 屈文忠. Lamb波高斯混合模型螺栓松动损伤检测[J]. 机械科学与技术, 2020(4): 493-500.
WANG G, XIAO L, QU W Z. Lamb wave gaussian hybrid bolt loose damage detection[J]. Mechanical Science and Technology, 2020(4): 493-500.

[118] 李敏, 崔树芹, 谢治平. 高斯混合模型在印花织物疵点检测中的应用[J]. 纺织学报, 2015, 36(8): 94-98.
LI M, CUI S Q, XIE Z P. Application of Gaussian mixture model in defect detection of printed fabrics[J]. Journal of Textile Research, 2015, 36(8): 94-98.

**作者简介**

**陈世超**(1987- ), 男, 澳门科技大学计算机技术及应用专业博士生, 中国科学院自动化研究所复杂系统管理与控制国家重点实验室助理研究员, 主要研究方向为数据处理、工业物联网、边缘计算。

**崔春雨**(1998- ), 女, 就职于中国科学院自动化研究所复杂系统管理与控制国家重点实验室, 主要研究方向为数据处理、边缘计算。

**张华**(1986- ), 女, 博士, 北京航天智造科技发展有限公司平台研发部高级工程师, 主要研究方向为现代精密测量、工业物联网和边缘计算。

**马戈**(1990- ), 男, 博士, 中国工业互联网研究院智能化所工程师, 主要研究方向为工业互联网、人工智能、边缘计算等。

**朱凤华**（1976- ），男，博士，中国科学院自动化研究所复杂系统管理与控制国家重点实验室高级工程师，主要研究方向为人工交通系统、平行交通管理系统。

**商秀芹**（1983- ），女，博士，中国科学院自动化研究所复杂系统管理与控制国家重点实验室助理研究员，主要研究方向为智能制造的数据驱动建模与优化技术。

**熊刚**（1969- ），男，博士，中国科学院自动化研究所复杂系统管理与控制国家重点实验室研究员，主要研究方向为复杂系统平行控制与管理、智能制造、智能交通。

**收稿日期：**2020-03-16

**通信作者：**熊刚，xionggang@casc.ac.cn

**基金项目：**国家重点研发计划基金资助项目（No.2018YFB1702701）；国家自然科学基金资助项目（No.61773381，No.U1909204，No.U1811463，No.61773382，No.61533019）；北京高等学校高水平人才交叉培养实培计划“智能边缘计算技术与设备”项目；东莞创新领军人才项目（熊刚）

**Foundation Items:** The National Key Research and Development Program of China (No.2018YFB1702701), The National Natural Science Foundation of China (No.61773381, No.U1909204, No.U1811463, No.61773382, No.61533019), Practical Training Plan Project “Intelligent Edge Computing Technology and Device” of High-level Talents Cross-cultivation in Beijing Universities and Colleges, Dongguan’s Innovation Talents Project (XIONG Gang)

# 基于分层注意力网络的方面情感分析

**宋婷[1]，陈战伟[2]，杨海峰[1]**
1. 太原科技大学计算机科学与技术学院，山西 太原 030024；
2. 中国移动通信集团山西有限公司，山西 太原 030001

**摘要**
基于深度学习的方面情感分析是自然语言处理的热点之一。针对方面情感，提出基于方面情感分析的深度分层注意力网络模型。该模型通过区域卷积神经网络保留文本局部特征和不同句子时序关系，利用改进的分层长短期记忆网络（LSTM）获取句子内部和句子间的情感特征。其中，针对LSTM添加了特定方面信息，并设计了一个动态控制链，改进了传统的LSTM。在SemEval 2014的两个数据集和Twitter数据集上进行对比实验得出，相比传统模型，提出的模型的情感分类准确率提高了3%左右。

**关键词**
深度学习；方面情感；区域卷积神经网络；分层长短期记忆网络；注意力机制；动态控制链

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2020045

# *Aspect sentiment analysis based on a hierarchical attention network*

**SONG Ting[1], CHEN Zhanwei[2], YANG Haifeng[1]**
1. College of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China
2. China Mobile Communications Group Shanxi Co., Ltd., Taiyuan 030001, China

***Abstract***
Aspect sentiment analysis based on deep learning is one of the hot spots in natural language processing. Aiming at aspect sentiment, a deep hierarchical attention network model based on aspect sentiment analysis was proposed. The local features of the text and the temporal relationship of different sentences were retained in model through the convolutional neural network, and the emotional features within and between sentences were obtained by using the layered long short-term memory network (LSTM). Among them, specific aspects of information were added to LSTM and a dynamic control chain was designed to improve the traditional LSTM. A comparative experiment is conducted on the two data sets in SemEval 2014 and the Twitter data set. Compared with the traditional model, the accuracy of sentiment classification of the proposed model increases by about 3%.

***Key words***
deep learning, aspect sentiment, regional convolutional neural network, layered long short-term memory, attention mechanism, dynamic control chain