特约专栏

材料数据库和数据挖掘技术的应用现状

杨 丽[1,3]，苏 航[1,2]，柴 锋[1]，罗小兵[1]，段琳娜[1,2]

（1. 钢铁研究总院工程用钢研究所，北京 100081）

（2. 北京钢研新材料科技有限公司，北京 100081）

（3. Institute of Iron and Steel，RWTH Aachen University，Aachen 52072，Germany）

摘 要：数据库作为材料基因工程的三大重点研究领域之一，对材料的加速设计有重要的作用和意义。从 20 世纪 70 年代起材料数据库开始在国内外逐步发展，到目前为止各国都已形成了一定数量的离线和在线数据库。早期材料数据库的基本功能是数据存储、数据管理以及数据检索服务。随着材料基因工程理念的提出，材料数据库开始关注发展数据共享、数据自动收集和输出等系列新功能。通过在线集成计算软件或程序、数据分析工具已逐步建立了一些基于数据库的材料智能设计平台。数据挖掘不考虑复杂的物理和化学意义，而是直接从材料数据库中挖掘出有价值的知识或模式的过程，它能够充分发挥材料数据库甚至小数据量在材料设计中的作用。目前已经被广泛应用到材料性能预测和优化、缺陷质量预测和生产监控、微观组织识别等相关领域。未来，材料数据库和数据挖掘技术将更加紧密地结合，协同向材料集成计算和智能设计的需求方向发展。

关键词：材料数据库；材料基因工程；数据挖掘；集成计算；性能预测

中图分类号：TB30 **文献标识码**：A **文章编号**：1674-3962(2019)07-0672-10

Material Database and Application Status of Data Mining Technology

YANG Li[1,3], SU Hang[1,2], CHAI Feng[1], LUO Xiaobing[1], DUAN Linna[1,2]

(1. Department of Structural Steel, CISRI, Beijing 100081, China)

(2. Beijing CISRI New Material Technology Co., Ltd., Beijing 100081, China)

(3. Institute of Iron and Steel, RWTH Aachen University, Aachen 52072, Germany)

Abstract: Database as one important research direction of materials genome initiative (MGI) has a meaningful role in the accelerated design of material. The material database was gradually developed worldwide since 1970s, and then a huge number of offline and online databases have been built up until now on. The basic function of early database is to store, manage and search data. With the concept of MGI put forward, material database begins to focus on a lot of advanced functions, such as data sharing, automatic collection and data output. Intelligent design platform for material based on the material database has been developed by online integrated computing software or program and data analysis software. Data mining is a process to find the valuable knowledge from the database without considering the complex mechanism of physics and chemistry. It plays an important role in material design in database or small data set. Data mining has been widely applied to the prediction and optimization of material properties, defects or quality prediction and production monitoring, recognition of microstructures. In future, the material database and data mining technology will be integrated tightly to satisfy the requirements of the integrated computing and intelligent material design.

Key words: material database; materials genome initiative; data mining; integrated computing; properties prediction

1 前 言

20 世纪 60 年代 IBM 数据库管理产品 IMS 技术的推出，为数据库的发展奠定了基础。随后，各国先后开始建立材料数据库，为材料标准、科研数据提供结构化的储存途径以及信息查询等功能。

2011 年美国提出发展材料基因工程，即数据库、高

收稿日期：2018-12-03 **修回日期**：2019-01-21

基金项目：国家重点研发计划项目(2017YFB0703002，2017YFB0701801)

第一作者：杨 丽，女，1986 年生，工程师

通讯作者：苏 航，男，1969 年生，教授级高级工程师，博士生导师，Email：hangsu@vip.sina.com

DOI：10.7502/j.issn.1674-3962.201812002

通量计算方法与高通量实验方法三大要素，为加速材料的智能设计作技术支撑。材料数据库的作用和地位随之变得更加突出：一方面，材料数据库可为高通量实验以及高通量计算结果提供海量数据存储空间；另一方面，材料数据库为高通量计算提供参数，或通过挖掘数据库中的知识模型，指导材料设计。

数据挖掘是数据库发现知识模型的重要方法，是一个通过从不完全的、有噪声的、模糊的、随机的大型数据库中，发现隐含的、未知的、可能有用的并且最终能被理解的模式的重要过程。虽然早在 20 世纪初期基于数据挖掘的数学基础就已基本成熟，但直到计算机的出现和计算能力的提升，大数据分析、数据挖掘等操作才变得更加切实可行。将数据挖掘方法应用到材料数据库的规律学习中，是指导新材料设计开发的一个重要手段。

本文针对国内外材料数据库和数据库技术的发展应用现状进行了综述，根据材料研发和理性设计新模式的发展需求，讨论了构建材料基因工程所需的材料数据库和数据挖掘技术目前存在的问题和未来发展方向。

2　材料数据库

2.1　传统材料数据库

以欧美、日韩等为代表的发达和新兴工业国家从 20 世纪七八十年代起，先后开始发展材料数据库，目前都已拥有一定数量的材料数据库，涵盖了黑色金属、有色金属、高温材料、复合材料、陶瓷材料、橡胶、核工业材料、功能材料等各种材料的成分、相图、晶体结构、性能参数等数据[1-3]。我国也从 20 世纪 80 年代开始由科研院所、企业自主建立了大量不同规模、分散独立的材料数据库，如钢铁研究总院的合金钢数据库、中国航发北京航空材料研究院的航空材料数据库、北京有色金属研究总院的有色金属数据库、清华大学的新材料数据库、西北工业大学的复合材料数据库、北京机电研究所的材料热处理数据库等上百个专业材料的数据库[4]。

根据存储数据种类的不同，材料数据库主要分为：材料热力学和相图数据库、晶体结构数据库(如无机晶体学数据库(ICSD))、材料性能数据库(标准或实验)、工艺性能数据库(如热处理数据库、金属切削数据库等)、特殊性能数据库(如腐蚀数据库和疲劳数据库)、专用数据库(如航空材料数据库、汽车材料数据库)等。根据存储数据形式的不同，数据库可分为数值型、文献型和文献/数值综合型。根据存储数据的服务模式，可分为离线型数据库和在线型数据库。由于早期建立的传统材料数据库主要是离线型，多服务于研究机构或组织的数据存储和研究，存在规模小、用户局限性高、商业化程度不高等缺点，因而其更新和应用受到人力、物力的限制，甚至部分数据库逐渐销声匿迹。

随着 web 网络技术的普及和快速发展，国内外较活跃的材料科学数据库开始以在线方式管理和服务，提高了材料数据库的商业化程度，强化了对用户的服务膜式。在线数据库的主要优势是更易推广和数据共享，通过将数据库商品化为外部机构提供有偿服务，间接推动了数据库的应用和全面快速发展。目前，国际知名的商业化材料在线数据库有美国的 MatWeb 和 ASM International、瑞士的 Total Materia、日本的 NIMS、德国的 Key to Steel 等，详情如表 1 所示[5]。

表 1　国际知名在线材料科学数据库[5]

Table 1　International famous online databases of material science[5]

Database	Country	Setup time	Field/Content
MatWeb	USA	Mid-1990s	A searchable material properties database, including ca. 59 000 materials: thermoplastic and thermoset polymers, metals (aluminum, cobalt, copper, lead, magnesium, nickel, steel, superalloys, titanium and zinc alloys), ceramics, semiconductors, fibers and other engineering materials
Total Materia	Switzerland (Key to Metals AG)	1999	One largest online database for metallic material properties, including ca. 15 million property records for over 350 000 metallic and also non-metallic materials
NIMS	Japan (NIMS)	2001	One largest online material database, including 17 databases: structural materials database (creep, fatigue, corrosion, strength, microstructure data), engineering database (CCT diagram, *etc.*), and database for superalloys, non-metallic materials, and physical properties (phase diagram, diffusion data, *etc.*)
Key to Steel	Germany (Verlag Stahlschlüssel GmbH)	2002	Database only for steel, including more than 70 000 standards and steel brands of approx. 300 steelworks and suppliers, and providing the steel searching service based on designation, composition and properties
ASM International	USA (ASM)	2002	Including database for alloy, phase diagram, failure analysis, micrograph, corrosion analysis and medical material, *etc.*

我国材料数据库的商业化发展也随着移动互联网的兴起得到极大提速。以钢研·新材道、材易通、欧冶知钢为代表的一批在线数据库服务平台先后出现。其中钢研·新材道的“全球钢材高端云服务”是依托于钢铁研究总院国内顶尖研发团队和 65 年的技术积淀建立起来的材料大数据和云服务平台，其 Atsteel 在线材料数据库包含上千个国内外标准、上万个牌号的材料性能数据，以材料大数据和定制研发为核心理念，致力于技术市场化的“互联网+”之路，为中高端材料用户提供研、产、检、造、用的全产业链服务。成都材智科技有限公司建立的 MatAI 材料智能设计平台具有能够根据用户需求提供数据管理和新材料设计优化等新功能。

传统材料数据库的主要功能是数据存储和数据管理，同时还提供数据检索服务，方便用户快速获取感兴趣的数据信息。例如日本的 NIMS 数据库就专门配套建立了 MatNavi 检索系统，使用户可以根据关键字/数值、树形节点对数据库的相关内容进行检索。美国 MatWeb 数据库也提供了基于数值、关键内容、类别的检索方法。我国钢研·新材道的 Atsteel 在线材料数据库增强了数据库的检索功能，除了以关键字、材料牌号检索的方式外，还提供成分、性能的区间范围值及其他多参数组合的高级检索功能，满足用户的各种检索需求。

2.2 材料基因工程的共享数据库

美国提出的材料基因工程理念，形成了材料数据库的新发展方向。目前，欧美国家建立材料基因工程数据库，除了发展新学科的独立材料数据库外，更希望搭建一个包含各种硬件、软件和专用数据传输标准的数据共享平台，如美国正在建设的 Globus 数据库平台[6]。通过特殊的信息工程技术，保证大数据易存储和搜寻等功能，既可将各地分散的传统材料数据库连入整个材料基因数据库共享平台，又可鼓励科研人员上传、发布新的科学成果，共享数据集；通过合理的材料数据库传输标准设计，满足各学科的数据存储需求和应用；而且通过数据库平台的软件集成进行在线计算，实现数据自动收集和数据挖掘，如 Material Project 平台。

促进材料基因工程数据库建设和发展的关键是数据共享。美国在数据共享方面采取了很多措施，21 世纪初期为了促进“人类基因组”项目数据库的建立，鼓励科学家快速分享 DNA 数据，提倡在 24 h 内上传到公共 GenBank 数据库中[7]。随着材料基因工程理念的提出，美国科学技术政策局(OSTP)和美国国际开发署(USAID)于 2013 年和 2016 年先后出台了“公共访问计划”，要求由 OSTP 和 USAID 等资助的科学研究数据需要在一定时间内公开，使公众、企业和其他科学人员能够获取[8]。美国国家科学基金委(NSF)也推出了“宣传和共享研究结果”的政策，鼓励科学人员能够共享在 NSF 资助的工作过程中创建或收集的主要数据、样本、实物和其他材料[9]。我国的科学数据共享工程自 2001 年底启动了气象科学数据共享试点以来，已在 24 个部门开展了相应的科学数据共享工作。整体而言，目前国内外的数据共享工作，主要是先通过科研联盟进行再不断扩散，并建立数据贡献积分制度显示不同科研用户的数据贡献率，从而间接反映其在相关领域的成果和影响力。

为了保护共享数据的权利和所属，目前国内外的共享数据库平台借鉴期刊论文模式，为每个上传的科学数据(集)注册唯一的 DOI 标识符，促进数据的保存、参考和引用[10]。美国材料数据平台(MDF)建立的可以发布数据以及查询数据的共享数据库平台 Globus，就是基于 DOI 对数据进行标识。通过该平台，可以搜索 MDF 连接的各种数据库/数据集里面保存的所有计算和实验数据，包括 NanoMine、PPPDB、Khazana Polymers、Khazana VASP、JANAF、SLUCHI(VASP)、Crystallography Open Database、Classical Interatomic Potentials、XAFS Data Library、OQMD 等十几个数据库。我国也积极推动共享数据库、在线数据库的发展，搭建了“材料科学数据共享网”平台，集合了分布在全国各地的 30 余家科研单位的海量数据资源，包括黑色金属、有色金属、复合金属、有机高分子、无机非金属等各类材料科学数据，为国家基础条件建设提供了雄厚的材料科学数据资源共享服务与应用支撑[11]。该平台目前也是通过提供标准的数据 DOI 注册系统以及数据采集标准，保证上传数据的标识性和结构化。近年来，随着区块链技术的不断成熟和发展，已有一些将区块链技术引入到材料数据库中的设想，实现对数据来源的标记，进行数据的版权保护，激发大家共享数据的热情。

高质量的共享材料数据对于材料基因工程具有重要的意义，不仅可以作为模拟计算的输入参数，也可以作为知识发现的样本数据，还可以为发现新的理论和技术提供线索。因此，数据的可信度是构建材料数据库时需要关注的一个重要问题。目前的主要解决方法是：一方面通过领域专家或数据库专员进行数据审核，并提供领域专家认证码，保证数据的可信度；另一方面建立完整规范的统计数据质量控制体系，通过进行相似数据的对比，判断数据的可信度或进行数据补充和修复[12]。

2.3 材料基因工程数据库的发展方向

除了数据共享、存储和查询外，材料基因工程的数据库还需要加强对分散的、已建立的数据库进行整合、利用，通过软件集成实现数据自动收集功能，为大数据

的学习和数据挖掘提供数据，指导新材料的研发。因此，材料基因工程的材料数据库开始发展如数据库匹配、数据自动收集、在线可视化、在线集成计算、在线分析等新功能。

2.3.1　数据库的匹配功能

数据库的自动匹配技术是将人工智能技术、模式识别等数据挖掘方法应用到材料数据库中，建立数据库之间的数据关联性，是数据挖掘技术在材料数据库中的一个成功应用。在数据库“云”概念的基础上，通过数据库的自动匹配算法可以实现“云”中的分布式数据库、异构数据库或多类型文件之间的连接，如图 1 所示[13]。

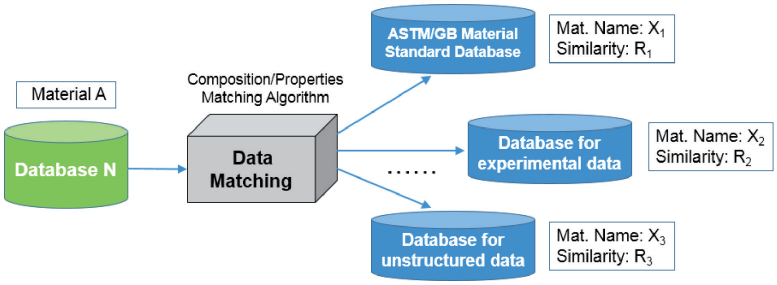

图 1　数据库匹配技术流程图[13]

Fig. 1　Flow chart of data matching procedure[13]

数据库自动匹配功能的主要优势是可以解决不同材料数据库之间存在的数据结构差异性、各国材料标准牌号和命名方式的不一致性、数据上传文件格式的多样性以及单一数据库中的信息不完整性等问题。在材料数据库中使用数据自动匹配技术，可以实现“小数据”到整个数据库系统的关联，获取相近材料的完整性能数据，是“小数据”换“大数据”的共享过程，也是实现分散数据库之间关联的一个重要方法。

德国的 Key to Steel 以及 Matmatch 等部分商业化在线数据库具有一定的多国牌号对照匹配查询以及数据库中相似材料的查询功能，但应用范围比较窄，仅适用于国内外产品牌号数据信息的对比。而我国的 Atsteel 数据库配套开发了多国钢铁材料牌号的自动匹配技术和功能，既可以实现各国相似材料牌号之间的关联匹配，还可以实现标准数据库、实验数据库、私有数据库等不同数据库之间的关联查询。目前该项数据匹配技术已经推广到钢铁材料的焊材匹配应用中，可以为焊接母材与焊材的匹配提供合适的材料选择方案。以 460 MPa 强度级别的系列钢材为例，基于北京钢研新材科技有限公司的钢铁数据库和焊接数据库，利用数据匹配技术进行了母材和焊材的匹配设计，如表 2 所示。可见通过数据匹配技术为母材设计匹配的焊材，基本与《焊材手册》推荐的相同强度级别的材料相吻合。其中，由于新的焊材数据库包含了最新的焊材牌号，因而数据匹配算法给出的很多结果是一些新的焊材牌号。目前国外还没有见到有任何关于母材-焊材匹配计算的相关报道，而且国外的焊材数据库也较少，大多为焊接工艺数据库。

瑞士 Total Materia 数据库开发的 SmartComp 材料智能判断功能相当于一种匹配检索功能，主要是通过对来自光谱仪或其他分析来源获得的金属化学成分进行智能识别，获得对应的材料金属牌号，为材料的智能识别和数据库自动分辨数据提供了新思路和方向。

表 2　基于数据匹配的母材-焊材匹配计算结果

Table 2　The matching results of base metal and welding material with data matching method

Recommended by ‘welding material handbook’		Recommended by the data matching method	
Base metal	Welding material	Industrail type	GB material
Q460E Thickness: 16~40 mm, $R_{p0.2}\geqslant$ 440 MPa, Impact T: T=−40 ℃	E5515-G E5516-G E5915-G E6215-G	CHE557Ni	E5515-GP
		THJ557RH	E5515-G
		GEL-557	E5515-GAP
		THJ556RH	E5516-G
		JQ.J557RH	E5515-GP
		J557RH	E5515-N1P
		J556RH	E5516-N1P
		J607RH	E5915-G
		CJ557HG	E5515-G
		J607Ni	E6215-G

2.3.2　数据库的数据收集和输出功能

数据的收集功能决定了数据库的发展规模和活力。建立数据的自动收集和输出功能，实现数据库与高通量实验、高通量计算的连接，是材料基因工程数据库发展的另一个重要方向。

互联网、云数据技术的发展在一定程度上为数据的收集、积累提供了支撑。共享数据库通过提供数据自主上传的接口，可实现用户自服务的数据收集上传功能。

国家材料环境腐蚀平台建立了“腐蚀大数据”和环境数据的大通量高密度采集、无线传输及入库的功能，可实现数据库数据的自动积累。目前国内外团队开始研究新型软件，可自动通过阅读材料科学实验论文获取晶体结构等相关信息，为数据的自动收集提供了便利[14]。但是如何通过论文信息的数字化识别全面获取数据、数据来源及实验条件，也是需要考虑的一个重要问题。

面对用户对数据库的输出需求，目前一些在线数据库可根据用户权限有针对性地为用户进行数据分析、建模计算从而提供相关数据及格式的输出功能。MatWeb 数据库就为用户提供以 CSV、Excel 等格式输出数据库中数据的服务，方便用户线下对数据进行对比分析。此外，还提供输出包含材料参数的通用计算软件专用格式文件，可直接应用于 Solidworks、ANSYS、COMSOL 等软件的结构材料计算建模中。

2.3.3 数据库的在线集成计算和分析功能

材料基因工程数据库的另一个重要发展方向是能够在数据库的基础上实现在线分析、软件集成计算以及数据结果自动存储等功能。

通过在线集成第一性原理、热动力学等成熟的材料计算软件或程序进行计算，能够为数据库补充大量的材料结构、性能、相变等特征参量，而计算获得的数据同样能够用于数据挖掘和指导新材料的开发。在材料基因工程计划中，美国能源部(DOE)牵头伯克利实验室负责建立的 Material Project 就是一个数据库集成平台，其包含了 600 000 多种材料和数据，提供了第一性原理的材料计算平台，允许用户对计算数据进行共享，目前已有超过 20 000 名用户利用该平台进行新材料设计和优化。杜克大学创建的 AFLOWlib 数据库，利用 AFLOW 材料高通量计算算法，通过在线集成 VASP、ESPRESSO 等软件，实现了对已知材料电子分布、晶体结构、能量计算以及新型材料结构的自动预测，并可自动存储计算结果到数据库体系中，通过高通量计算不断扩充数据库的数据量[15]。目前该数据库已有 10^6 数量级的不同材料，其中有超过 10^8 数量级的材料性能数据是通过计算获得的。美国西北大学推出的开放量子材料数据库(OQMD)、中国的 MatCloud 高通量材料集成设计平台也具有相似的工作机制，通过调用 VASP 或 CASTEP 等第一性原理软件在超级计算机上进行大批量计算，再将相应的计算结果保存到数据库中，最终通过大数据分析来指导新材料设计[16]。日本 NIMS 开发的 COMPOThermo 在线计算软件，通过集成界面热导率数据库，可制定特殊热性能要求的复合材料。目前材料数据库集成第一性原理计算软件主要在功能材料的设计领域获得了较多成功的应用，同时在复杂的结构材料设计方面也有一定的应用。

此外，材料数据库也开始考虑数据的在线可视化、在线分析等功能。成都材智科技有限公司建立的 MatAI 材料数据管理平台可根据需求建立集成基础的数据对比分析、数据统计和可视化工具的材料数据库，以便在线进行散点图的分析、曲线的对比和统计的可视化。目前，一些数据库还可通过对热力学计算软件的集成连接，利用获得的材料热力学数据，配合数据库中其他数据共同进行数据挖掘和分析[17]。

3 数据挖掘方法在材料科学中的应用

3.1 数据挖掘方法简介

数据挖掘基本流程为：确定目标→数据库取样→数据预处理→数据挖掘建模→知识获取和解析→应用，如图 2 所示[18]。将清洗预处理后的样本数据分为 3 类：训练型数据、验证型数据和测试型数据，再用于模型学习、验证和测试。

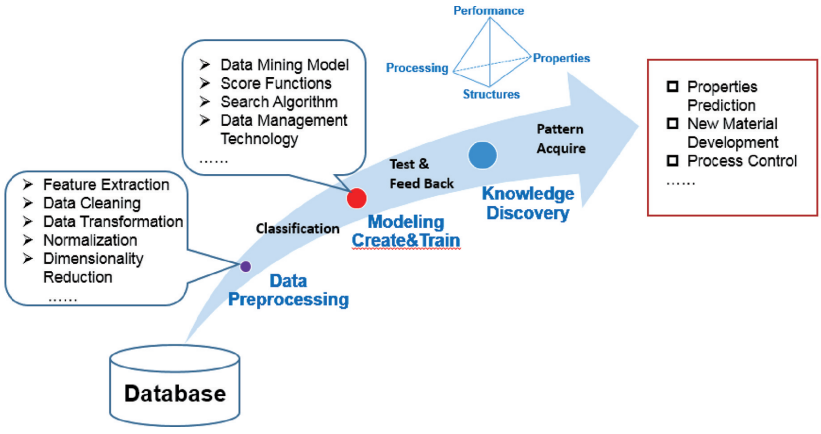

图 2 数据挖掘的基本流程[18]

Fig. 2 Flow chart for data mining procedure[18]

一个完整的数据挖掘算法通常是由模型结构、评分函数、搜索方法、数据管理技术几个基本模块组合构成[19]。例如一个反向传播神经网络(BP-ANN)数据挖掘算法通常是由神经网络模型结构、误差平方函数、参数梯度下降寻优等模块构成。组合不同的模型结构、评分函数、搜索方法等可以生成数量庞大的挖掘算法。此外，降维方法也被应用到数据处理中，如主成分分析(PCA)法就常被用于微观组织形貌等的降维处理，使得微观组织能够作为输入变量参与数据挖掘学习，从而通过回归、神经网络或其他模型方法最终建立工艺-微观结构-性能关系[20]。

数据挖掘的方法根据任务目的可分为预测性和描述性方法，根据学习方式可分为监督学习和无监督学习方法。在材料科学领域，目前常用的数据挖掘算法主要有：回归、分类、聚类、智能优化，如图 3 所示[21]。其中，神经网络和支持向量机是机器学习的两大主要流派，既可用于回归又可用于分类和优化。

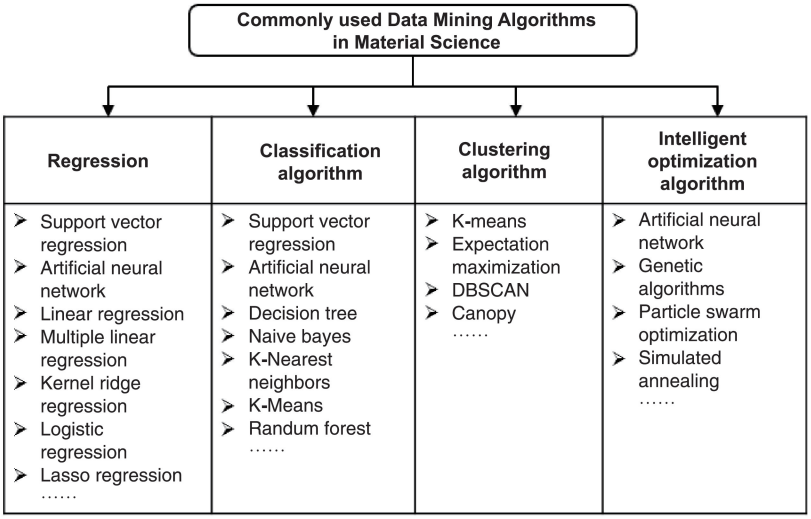

图 3　材料科学中常用的数据挖掘算法[21]

Fig. 3　The data mining algorithms used in material science[21]

神经网络最初起源于 1957 年 Rosenblatt 发明的单层感知机，随着非线性问题需求的增加，多层神经网络不断发展。神经网络基本原理是利用权重连接输入层、隐藏层、输出层之间的组合神经单元，并不断训练连接的权值直至计算结果足够逼近预期值，从而解决复杂的计算问题。随着多层神经网络的发展应用，深度学习的概念被提出，卷积神经网络、解积神经网络等更复杂的神经网络算法也随之出现，如图 4 所示[22]。

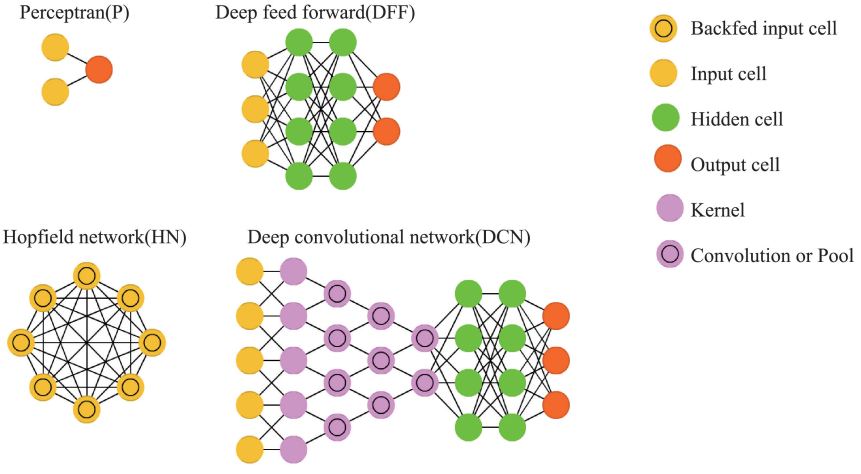

图 4　不同类型的多层神经网络[22]

Fig. 4　Various multi-layers neural networks[22]

支持向量机(SVM)是由 Cortes 和 Vapnik 等于 1995 年首先提出的，属于二分类模型算法，其基本原理是通过线或者超平面实现样本集在二维或三维空间里面的间隔最大化[23]。相较于其他分类统计算法对大样本数据量的要求和难以解决复杂的高维度问题，SVM 在解决小样本、非线性及高维度的数据模式识别时也能获得较好的结果，表现出了许多特有的优势，并能够被推广应用到函数拟合等其他机器学习问题中。

3.2 数据挖掘方法在材料科学中的应用

随着大数据的发展和计算机软硬件实力的提高，90 年代末期数据挖掘方法就已经开始被大量应用到材料科学研究及生产控制过程中，如材料性能预测和优化、新材料设计开发、生产过程的监控等方面。

3.2.1 材料性能预测和优化

数据挖掘在材料性能预测和优化方面的应用最为广泛。其中多层神经网络算法是使用较多的一种数据挖掘算法，常配合不同的优化算法进行解的快速搜索，如非线性最小二乘法、批梯度下降算法、冲量批梯度下降法、遗传算法等。常规性能预测算法基本思路是：假定已知某材料的一组性能指标 P 与 X 个因子之间的相关性，利用数据库中 n 个样本的实验数据集，设置各因子的可变范围以及约束条件，通过数据挖掘的方法，建立 P 与 X 之间的线性或非线性关系，并据此指导材料的单一或多目标优化。目前，数据挖掘在材料的强度[24]、冲击韧性[25]、淬透性[26]、疲劳和蠕变[27]等相关性能预测方面已有大量的应用。

基于热轧钢板的成分、热轧工艺(温度、变形、道次)等实际数据，Yang 等[28]通过 3 层前馈神经网络模型，结合贝叶斯对权值进行优化训练的方法，获得了误差较小的拉伸强度预测结果。Powar 等[29]通过 11-5-7 的 3 层神经网络结构，建立了包含 30CrMoNiV5-11 的元素成分、奥氏体化温度和时间、冷却时间 $t_{8/5}$ 等的输入层，与由屈服强度、抗拉强度、伸长率以及珠光体、贝氏体和残余奥氏体的体积分数等构成的输出层之间的关系模型，且相关性系数 R 大于 90%。针对相变诱导塑性(TRIP)钢，Bhattacharyya 等[30]利用 11-15-1 的 3 层神经网络模型，采用双曲正切函数作为传递函数，获得了包含 C，Si，Mn，P，Al，Nb，Cr 的质量分数、临界区退火温度和时间、贝氏体等温转变温度和时间的 11 个输入层节点到残余奥氏体含量的预测模型。Liu 等[31, 32]利用前馈神经网络模型对 Nb-Si 基高温合金的微观组织与性能之间的关联关系进行了挖掘学习，建立了基于 Nb_5Si_3 的体积分数、形貌、尺度等微观组织变量对抗拉强度、断裂韧性等实现预测的模型。

遗传算法-神经网络(GA-ANN)结合算法被应用到了某 FeCrNiMn 奥氏体不锈钢体积模量的预测中，且该预测结果与基于密度泛函理论(DFT)的第一性原理的计算结果非常接近，证明了 GA-ANN 算法预测的精准性[33]。此外，在已获得的第一性原理计算结果数据基础上利用随机森林等方法构建数据挖掘模型，获取知识模型和重要的影响因素后，即可代替第一性原理计算直接预测 Ni 基、Co 基高温合金掺杂元素的置换能和几何结构，间接节约了材料性能计算和设计的时间[34]。可见，数据挖掘为第一性原理计算的加速提供了另一种思路和方向。

3.2.2 材料特征曲线拟合

数据挖掘算法在材料特征曲线的拟合方面也有着广泛的应用。Haque 等[27]利用神经网络，对获得的大量实验数据进行拟合，建立了不同马氏体含量的系列双相钢的腐蚀疲劳裂纹扩展速率 da/dN 与应力强度因子变化量 ΔK 的关系模型，实现了其在双相钢腐蚀疲劳裂纹扩展速率预测中的应用。

在热塑性变形方面，通过对材料流变应力应变实验数据的学习，针对不同材料成分，可拟合和预测应变速率和温度条件下对应的高温热压缩时的流变应力应变曲线和本构方程，以及动态再结晶的体积分数和晶粒尺寸，从而为后期锻造过程的多场耦合建模、应力应变计算和组织预测模拟提供精准的材料本构方程[35]。然而，利用数据挖掘的模型分析成分对流变应力的影响还有待进一步深入的研究。

在焊接方面，数据挖掘算法除了被应用到材料焊接后的性能预测(如热影响区的硬度[36])，还被应用到了焊接热源形状参数的拟合预测中。例如通过对实际钨极惰性气体保护焊接(GTAW)过程中获得的不同焊接条件(如电流、焊接速度)下双椭圆体热源尺寸数据集进行数据挖掘，可较好地拟合出焊接热源形状参数变化情况，并预测未知焊接条件下的形状结果[37]。通过拟合预测热源模型，能够为焊接过程的有限元模拟提供精准的热源输入模型，保证了更准确的温度场计算结果。

3.2.3 质量预测及生产监控

基于风险最低原则，常采用支持向量机、决策树、神经网络等分类算法对材料生产过程参数进行在线异常监控以及质量预测。

在钢生产过程中的表面质量分类和缺陷在线预测控制方面，数据挖掘算法已经获得了较多的实际应用，基本上能保证预测和监控精准度在 90%以上[38]。其基本监控流程是：通过在线缺陷图像信息采集，获取缺陷图片的几何特征(如长度、正方度、面积等)、图片的灰度数据、织构特征信息(能量、粗糙度、对比度、方向等)等

表征参数，再利用数据挖掘中的分类算法和优化算法组合建模，快速实现缺陷的鉴定、识别和分类[39]。

分类算法还被广泛应用到焊接质量预测控制等相关方面。通过决策树分类模型，根据焊接过程中的电流和电压信号可以实现对焊接效果(有气孔、完好、过烧)的评价，对焊接效果等级进行分类和在线监控[40]；结合聚类和神经网络的数据挖掘算法，可基于数据库中焊接缺陷分类结果，判断影响焊接稳定性的因素[41]；利用支持向量机可对焊接的高热输入风险进行在线评估和预测[42]。

此外，对材料服役过程的缺陷诊断，也能够使用分类算法。决策树和支持向量机等就被应用到对滚动轴承缺陷的分类和诊断工作中，通过前期数据的学习和模型建立，使得根据轴承的震动信号就可自动实现对缺陷状况的诊断[43]。

3.2.4　微观组织的识别和分类

与指纹识别功能类似，数据挖掘方法也开始被应用到对材料微观组织照片的识别和分类中，使得组织信息能够数字化，为高通量实验或数据库的非结构化文件的分类和关联提供了新的思路和方向。

Decost 等[44]利用支持向量机算法实现了对黄铜、球墨铸铁、灰口铸铁、亚共析钢、高温合金、退火孪晶等不同系列微观组织照片的识别和分类，以便对存放有大量材料组织照片的数据库进行分类管理。此外，Gola 等[45]利用支持向量机算法也实现了对金相组织照片和透射电镜照片中出现的马氏体、贝氏体和珠光体的基体组织进行分类。

此外，数据挖掘方法以及 PCA 等降维方法也开始被应用到了三维场离子显微镜分析中，以获得更精准的数据结果[46]。PCA 主要是通过对数据进行特征值分析，确定出需要保留的主成分个数，舍弃其他数据冗余和噪声，从而实现数据的降维。PCA 是目前图像处理较为常用的降维方法。

3.3　数据挖掘在材料基因中的应用发展和问题

数据挖掘过程不需要考虑参数之间复杂的物理和化学意义，就可以直接从材料数据库中挖掘出有价值的知识或模式，它能够充分发挥材料数据库甚至小数据量在材料设计中的作用。在材料基因工程项目的推动下，数据挖掘在材料设计中的应用不断被深入和拓展。

根据材料基因工程理念，数据挖掘算法未来可以被集成、应用到材料数据库以及高通量计算平台中，通过对材料成分-工艺-组织-性能数据规律和知识的自动学习，进行多参数、多目标的优化计算，能够大大提高材料设计速度，降低设计成本，更好地指导材料性能预测或新材料设计。目前，基于材料数据库和高通量计算结果，数据挖掘技术已经开始成功运用到了功能材料等新材料的设计和开发中。徐一斌团队[47]在数据库基础上，通过支持向量机、回归等机器学习方法获得了高界面热阻的材料组合，并结合高通量薄膜制备技术，制备出了目前世界上隔热性能最高的无机纳米复合薄膜。

数据挖掘算法的复杂性以及材料数据库中相关参数的多样性，决定了数据分析是一个需要多学科知识交汇和大量经验积累的过程。Agrawal 等[48]基于 NIMS 数据库中的钢铁材料疲劳数据库，建立了针对材料疲劳强度设计的知识模型，对比了十几种数据挖掘组合算法的精准性，包括线性回归、决策树、支持向量机、人工神经网络、模型树等，并获得了包括材料成分、工艺参数、缺陷分布等 25 个输入参数对疲劳强度的正负相关性影响，如图 5 所示。因此，如何在已有材料数据库中确定自变

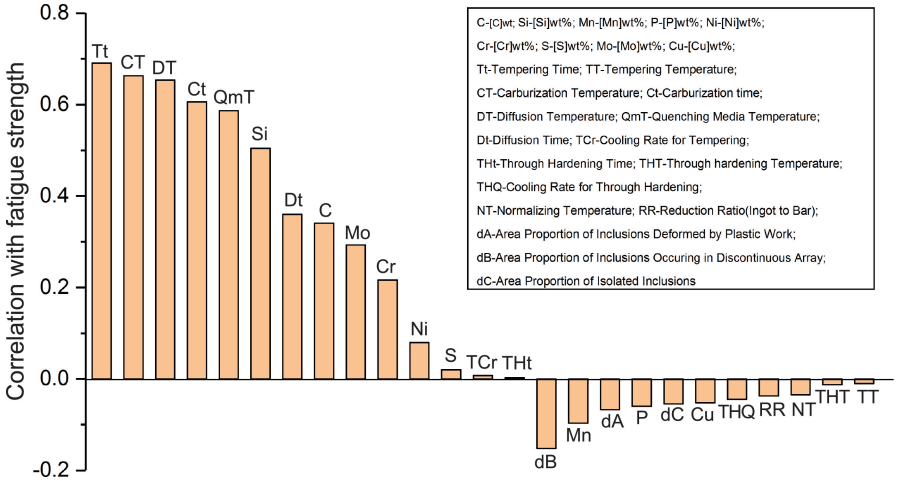

图 5　25 个不同参数与疲劳强度相关性的关系[48]

Fig. 5　The relationship between 25 inputs and correlation with fatigue strength[48]

量和因变量，并选择合适的数据挖掘算法，如何从获得的结果中读取知识，以及如何判断数据挖掘获得知识的准确性，是数据挖掘过程中需要深入研究的问题。

确保数据挖掘结果准确性的一个重要因素是材料数据库的数据可靠性。因此，在建立材料数据库的过程中通常要求设置数据审查机制，以保证数据库中所有上传数据的正确性。当然在数据挖掘过程中，通过数据预处理可以对噪声点、异常值进行清洗，一定程度上能够减小数据误差造成的分析结果偏差。然而，除了利用成功的实验数据进行数据挖掘和分析外，失败或不成功的实验数据用于预测新材料的合成也获得了较高的准确性[49]，大幅提高了新材料研发的可能性。

4 结 语

在材料基因工程中，数据挖掘需要与材料数据库以及高通量计算相互结合、协同发展，才能更好地发挥其对材料加速设计的作用和意义。

(1)数据库作为数据管理和存储技术，为数据挖掘和高通量计算提供了输入参数。材料数据库目前已逐步从孤立的离线数据库向在线数据库和共享数据库方向发展，但其结构化、标准化等方面还有待改善。逐步发展起来的数据库云理念结合数据匹配算法方便了分布式数据库之间的连接，为数据库结构差异性问题提供了解决途径。同时，需要进一步扩大数据量以实现材料数据库的规模化进而提高数据挖掘结果的精准性。

(2)数据挖掘可为材料数据库提供数据分析技术和方法，从已有的数据中发现知识和规律，加速材料设计。通过完善材料数据库中的材料成分、工艺、组织、性能数据，再利用数据挖掘技术可建立成分-工艺-组织-性能之间的关系模型。掌握从海量的数据中选择合适的样本数据、建立参数的相关性，并精准地提取规律和解释知识，是数据挖掘技术在材料设计中深入应用需要重点关心的方面。

(3)数据库与数据挖掘技术的结合、数据库匹配、数据自动收集、在线可视化、在线计算、在线分析等数据库新功能的拓展，将使材料基因工程数据库发展成为一个综合性平台，既是数据库平台，也是计算平台和数据分析平台。目前数据挖掘技术在数据库中的应用大多都是线下操作，而且数据样本的大小和数据的精准性也影响着数据挖掘的结果。未来，通过在材料基因数据库中直接集成嵌入数据挖掘算法，进行数据在线自动学习、异常数据清洗、知识提取，以便更好地支撑材料设计，提高研发效率。

参考文献 References

[1] RHYIM Y M, YIM C D, LEE C G, *et al*. Asian Materials Database Symposium[C]. Sanya: University of Science and Technology Beijing, 2010: 48.

[2] KIRCHHEINER R, KOWALSKI P, HARTUNG T. Materials and Corrosion[J], 1993, 44(10): 410-415.

[3] KARDITSAS P J, LLOYD G, WALTERS M, *et al*. Fusion Engineering and Design[J], 2006, 81(8-14): 1225-1229.

[4] 张大全, 吴崇田, 陆铖俊, 等. 上海电力学院学报[J], 2011, 27(5): 528-533.
ZHANG D Q, WU C T, LU C J, *et al*. Journal of Shanghai University of Electric Power[J], 2011, 27(5): 528-533.

[5] 李霞, 苏航, 陈晓玲, 等. 中国冶金[J], 2007, 17(6): 4-8.
LI X, SU H, CHEN X L, *et al*. China Metallurgy[J], 2007, 17(6): 4-8.

[6] National Science and Technology Council. Materials Genome Initiative for Global Competitiveness[R/OL]. (2011-06-24)[2018-12-03]. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/materials_genome_initiative-final.pdf.

[7] MARSHALL E. Science[J], 2001, 291(5507): 1192.

[8] USAID. Public Access Plan: Increasing Access to the Results of Federally Funded Scientific Research[R/OL]. (2018-03-08)[2019-01-21]. https://www.usaid.gov/open/public-access-plan/.

[9] NSF. Dissemination and Sharing of Research Results[R/OL]. (2019-01-07)[2019-01-21] https://www.nsf.gov/bfa/dias/policy/dmp.jsp/.

[10] DEHHARD I, WEICHSELGARTNER E, KRAMPEN G. Data Science Journal[J], 2013, 12: 172-180.

[11] 尹海清, 姜雪, 张瑞杰, 等. 中国科技资源导刊[J], 2016, 48(3): 65-71.
YIN H Q, JIANG X, ZHANG R J, *et al*. China Science & Technology Resources Review[J], 2016, 48(3): 65-71.

[12] CHUNG Y, KRISHNAN S, KRASKA T. Proceedings of the VLDB Endowment[J], 2017, 10(10): 1094-1105.

[13] 苏航, 张解, 陈晓玲, 等. 材料导报[J], 2005, 19(11): 8-11.
SU H, ZHANG J, CHEN X L, *et al*. Materials Review[J], 2005, 19(11): 8-11.

[14] SANGHOON P, BAEKJUN K, SIHOON C, *et al*. Journal of Chemical Information and Modeling[J], 2018, 58: 244-251.

[15] CALDERON C C, PLATA J J, TOHER C, *et al*. Computational Materials Science[J], 2015, 108: 233-238.

[16] 杨小渝, 王娟, 任杰, 等. 计算物理[J], 2017, 34(6): 697-704.
YANG X Y, WANG J, REN J, *et al*. Chinese Journal of Computational Physics[J], 2017, 34(6): 697-704.

[17] 王卓, 王礞, 雍歧龙, 等. 中国材料进展[J], 2017, 36(2): 132-140.
WANG Z, WANG M, YONG Q L, *et al*. Materials China[J], 2017, 36(2): 132-140.

[18] AGRAWAL A, CHOUDHARY A. APL Materials[J], 2016, 4(5): 1-17.

[19] 汉德, 曼尼拉, 史密斯. 数据挖掘原理[M]. 张银奎, 廖丽, 宋俊, 等译. 北京: 机械工业出版社, 2003.
HAND D, MANNILA H, SMYTH P. Principles of Data Mining[M]. Translated by ZHANG Y K, LIAO L, SONG J, *et al*. Beijing: China Machine Press, 2003.

[20] YABANSU Y C, STEINMETZ P, HÖTZER J, *et al*. Acta Materialia [J], 2017, 124(1): 182-194.

[21] LIU Y, ZHAO T, JU W, *et al*. Journal of Materiomics[J], 2017, 3 (3): 159-177.

[22] VEEN F V. The Neural Network Zoo[R/OL]. (2016-09-14) [2018-12-03]. http://www.asimovinstitute.org/neural-network-zoo/.

[23] CORTES C, VAPNIK V. Machine Learning[J], 1995, 20: 273-297.

[24] SUZUKI K, TOROGHINEJAD M R, ESFAHANI M B. Artificial Neural Networks-Industrial and Control Engineering Applications[M]. Croatia: InTech Press, 2011: 153-168.

[25] AZIMZADEGAN T, KHOEINI M, ETAAT M. Neural Computing and Application[J], 2013, 23(5): 1473-1480.

[26] TAGHIZADEH S, SAFARIAN A, JALALI S, *et al*. Materials & Design [J], 2013, 51: 530-535.

[27] HAQUE M E, SUDHAKAR K V. International Journal of Fatigue[J], 2001, 23(1), 1-4.

[28] YANG Y Y, MAHFOUF M, LINKENS D A, *et al*. Tensile Strength Prediction for Hot Rolled Steels by Bayesian Neural Network Model [C]. Proceeding of Automation in Mining, Mineral and Metal Processing. Chile: Curran Associates Press, 2009: 255-260.

[29] POWAR A, DATE P. Materials Science & Engineering A[J], 2015, 628(1): 89-97.

[30] BHATTACHARYYA T, SINGH S B, SIKDAR S, *et al*. Materials Science & Engineering A[J], 2013, 565: 148-157.

[31] LIU G, JIA L, KONG B, *et al*. Materials Science & Engineering A [J], 2017, 707: 452-458.

[32] LIU G, JIA L, KONG B, *et al*. Materials & Design[J], 2017, 129: 210-218.

[33] BENYELLOUL K, AOURAG H. Computational Materials Science[J], 2013, 77 (1): 330-334.

[34] 刘轶, 肖斌, 吴雨沁. 高通量计算和数据挖掘结合加速高温合金掺杂研究[C]//第二届材料基因工程高层论坛会议摘要集. 北京: 中国工程院出版社, 2018: 65-66.
LIU Y, XIAO B, WU Y Q. Accerlating the Doping Element Research of Supperalloys with High-throughput Computing and Data Mining Technology[C]// Proceeding of 2nd Forum of ICME. Beijing: Chinese Academy of Engineering Press, 2018: 65-66.

[35] KONG L X, HODGSON P D. ISIJ International[J], 1999, 39(10): 991-998.

[36] POURALIAKBAR H, KHALAJ M, NAZERFAKHARI M, *et al*. Journal of Iron & Steel Research International[J], 2015, 22(5): 446-450.

[37] TAFARROJ M M, KOLAHAN F. Fusion Engineering and Design[J], 2018, 131 (1): 111-118.

[38] LUIZ A O M, FLÁVIO L C P, PAULO E M A. Automatic Detection of Surface Defects on Rolled Steel Using Computer Vision and Artificial Neural Networks[C]// Proceeding of 36th Annual Conference on IEEE Industrial Electronics Society. USA: IEEE Press, 2010: 1081-1086.

[39] CUI D, XIA K. Mathematical Problems in Engineering[J], 2017, 16: 1-9.

[40] SUMESH A, NAIR B B, RAMESHKUMAR K, *et al*. Materials Today: Proceedings[J], 2018, 5(2): 8354-8363.

[41] ZHANG F, LUK T. IEEE Transactions on Electronics Packaging Manufacturing[J], 2007, 30(4): 299-305.

[42] CHEN J, WANG T, GAO X, *et al*. Computers in Industry[J], 2018, 94: 75-81.

[43] SUGUMARAN V, SABAREESH G R, RAMACHANDRAN K I. Expert Systems with Applications[J], 2008, 34(4): 3090-3098.

[44] DECOST B L, HOLM E A. Computational Materials Science[J], 2015, 110: 126-133.

[45] GOLA J, BRITZ D, STAUDT T, *et al*. Computational Materials Science[J], 2018, 148: 324-335.

[46] KATNAGALLU S, GAULT B, GRABOWSKI B, *et al*. Material Characterization[J], 2018, 146: 307-318.

[47] ZHAN T, FANG L, XU Y. Scientific Reports[J], 2017, 7(1): 1-8.

[48] AGRAWAL A, DESHPANDE P D, CECEN A, *et al*. Integrating Materials and Manufacturing Innovation[J], 2014, 3(1): 8.

[49] RACCUGLIA P, ELBERT K C, ADLER P D F, *et al*. Nature[J], 2016, 533: 73-76.

（编辑　王　瑶　吴　锐）