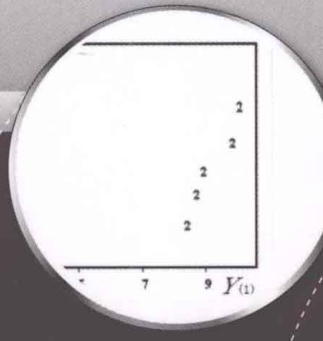
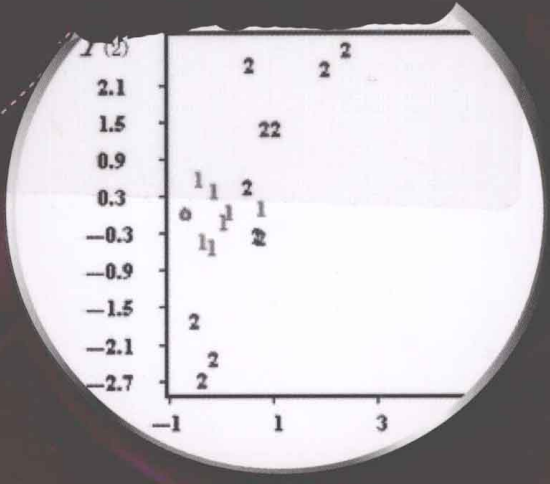
21 CENTURY
世纪计算机化学丛书

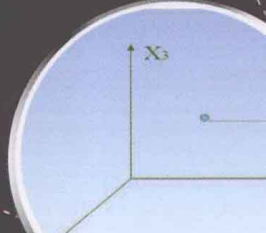
化学数据挖掘方法与应用

Chemical Data Mining and Applications

陆文聪　李国正　刘 亮　包新华　著







化学工业出版社

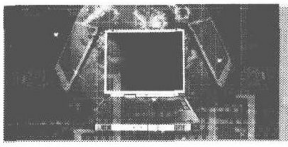
21 CENTURY
世纪计算机化学丛书

化学数据挖掘方法与应用

Chemical Data Mining and Applications

陆文聪　李国正　刘 亮　包新华　著

化学工业出版社
·北京·



21世纪
计算机化学丛书

序

计算机化学的兴起与发展是与化学知识创新的迫切需要紧密联系的。十年前化学家使用计算机的还不多，现在却已十分普及；十年前对化学计算的要求主要是化学信息的采集、加工、储存和利用，而如今除了以上的基本要求之外，更强调了由化学信息发现新知识和化合物物性的定量预测。计算机网络技术的飞速发展与普及，对计算机化学来说是一个发展的机遇，而愈来愈高的计算要求是计算机化学发展面临的新挑战。今天，以计算机及其网络深入到社会的各个层面为标志的数字化新世纪的到来，将使传统化学发生深刻的变化：以计算机及其网络系统为工具，建立由化学化工信息发现新知识和实现知识传播的理论和方法；认识物质、改造物质、创造新物质，认识反应、控制反应过程，创造新反应、新过程，将成为计算机化学研究的主体。化学数据挖掘、知识发现、计算机辅助结构解析、分子设计和合成路线设计等是当前计算机化学的主要研究方向。可以深信，在21世纪，数字化新世纪的化学不仅要靠“湿”实验室来发展，同时也要依赖于“干”实验室。所谓“干”化学实验室就是指数字化虚拟化学实验室。“干”、“湿”相结合才能更高效地孕育出新的化学实体，才能促进化学由实验科学向严密科学转化，才能大大提高化学非凡的创造力。

为了推广计算机化学的新理论、新技术和新方法，促进科技进步，我们策划了这套《21世纪计算机化学丛书》，主要介绍计算机化学近5年间的新理论、新技术和新方法。希望这套丛书不仅能够大大推动我国科技水平的进步，更能对我国生产力水平的提高产生巨大的影响。

陈凯先
2010年3月

前言

FOREWORD

计算机在理论化学和应用化学各个领域的广泛应用，极大地促进了化学学科的发展，并产生了一系列交叉学科，如计算（机）化学、化学计量学、化学信息学等。

化学化工领域积累了大量的科学实验和生产实际数据，如何总结这些数据中的规律性，进而用以指导以后的科学实验和生产操作，这是一项非常有意义的工作，这项工作的实施需要数据挖掘技术与化学化工知识和科学实践的结合。

所谓化学数据挖掘（Chemical Data Mining），就是利用机器学习方法对化学化工（或相关学科）中有关数据样本进行采集、整理、分析、建模等，试图归纳和总结数据中蕴含的规律性，进而利用所建定性或定量的数学模型预报未知样本的性质。化学数据挖掘的应用研究内容涉及材料设计、分子设计、化工过程优化等领域。化学数据挖掘方法和技术已成为化学信息学、生物信息学的主要研究工具。

利用化学数据挖掘方法和技术，可以总结药物分子的构效关系，即药物的生物活性与其结构特征参数（分子描述符）之间的定量或定性关系，在此基础上可以设计和预测新的高活性化合物。利用化学数据挖掘方法和技术，可以总结新材料的物理化学性质与其组成元素的原子参数、化学配方、制备工艺等参数之间的定性或定量关系，在此基础上可以辅助新材料研制和新产品开发，达到事半功倍的效果。利用化学数据挖掘方法和技术，对大型现代化工厂（特别是炼油厂、化工厂和炼钢厂）的生产操作过程作“工业诊断”，找出优化生产的“瓶颈”问题，建立解决“瓶颈”问题的数据挖掘模型，在此基础上可以实现低成本、高收率、低能耗、高质量地生产和制备各种化学产品。因此，利用化学数据挖掘所得研究对象的统计规律，可以指导我们更好地开展下一步的科学实验和生产实践，达到“事半功倍”的目的。化学数据挖掘方法和技术的应用成本低，却可能在科学实

验中节省人力物力，甚至在工业生产中产生可观的经济效益，因而化学数据挖掘方法和技术有广泛的应用背景。

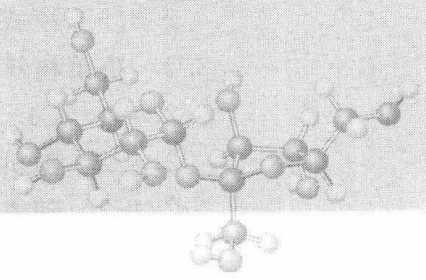
笔者长期从事化学数据挖掘方法在化学化工领域的应用研究工作，在该研究领域积累了大量成功应用实例，我们开发的化学数据挖掘软件 HyperMiner 和基于数据挖掘的工业优化控制系统已在国内若干大型企业得到实际应用，达到了增产降耗的目的。本书从化学工作者易于理解的角度介绍常用数据挖掘方法的基本原理，并重点介绍作者近年来在材料设计、工业优化、构效关系、生物信息学等领域的数据挖掘工作。

笔者曾与我国已故著名化学家陈念贻先生长期合作研究，很多工作曾得益于陈念贻先生的指导和帮助。笔者曾作为合作者协助陈念贻先生出版过两本学术专著，即《模式识别方法在化学化工中的应用》（科学出版社，2000）和《Support Vector Machine in Chemistry》（World Scientific Publishing Co. Pte. Ltd.，2004），本书的出版是笔者对于恩师陈念贻先生的化学数据挖掘工作在上海大学的继承和发展。本书有关科研工作得到了国家自然科学基金委员会、上海市科学技术委员会、上海宝山钢铁集团、云南省科技厅、北京石油化工设计院等单位的资助；有关学术研究和技术开发工作得到了笔者的研究生们的大力配合，其中刘旭和顾天鸿博士等在算法程序方面做了较多工作，杨善升和钮冰博士等在化学数据挖掘应用方面做了较多的工作；本书的出版得到了化学工业出版社的支持，在此一并致谢。

为方便读者学以致用，笔者为读者提供了化学数据挖掘应用软件 HyperMiner，读者下载后可免费使用 30 天（附录 1 含该软件简介和下载方法），希望广大读者能通过具体应用案例学习和受益。本书可供化学、化工及相关领域的科研人员和工程技术人员阅读，亦可作为高等学校的教学参考书。

化学数据挖掘涉及的研究领域很广，本书只是介绍了部分常用方法在笔者涉猎的研究领域中的工作，有关数据挖掘方法包括变量相关分析和多元统计、模式识别、人工神经网络、遗传算法、支持向量机、集成学习、特征筛选等；有关数据挖掘方法的综合应用案例涉及材料设计、工业优化、构效关系和生物信息学等领域。由于笔者的学识和工作所限，疏漏和不足之处在所难免，欢迎各位读者和研究同行提出宝贵意见。

陆文聪
2011 年 8 月于上海大学



目 录

CONTENTS

1 化学数据挖掘综述

1.1 化学数据挖掘的目的和意义

化学、化工是以实践为主的学科，其理论的发展往往落后于实践。认识物质、改造物质、创造新物质和认识反应、控制反应过程和创造新反应是化学、化工研究的主体。在长期的化学、化工实践中，人类积累了海量的化学、化工信息，这类信息散布在浩如烟海的各类化学、化工文献中，虽然这些化学信息为人们探索自然界的奥秘提供了基础，但由于数据量的迅猛增加却造成了使用上的困难，常规手段已无法满足化学、化工专家的需要，因此众多的化学、化工数据库应运而生。近年来，人们在利用数据库对化学、化工问题进行研究时，逐渐认识到海量数据的处理十分困难，有价值的规律性信息和知识还隐藏在数据内部。如何从化学、化工数据中发现更多、更有价值的化学、化工规律正逐步成为化学、化工专家关注的焦点，正如徐光宪先生在国家自然科学基金委员会成立十五周年庆祝大会上的讲话中所指出的那样[1]："从科学发展史看，科学数据的大量积累，往往导致重大科学规律的发现。……19 世纪 60 年代的化学积累了数十种元素和上万种化合物的数据，门捷列夫把这些元素按原子量的大小次序排序，发现它们化合物的性质有周期性变化，因而在 1869 年提出了元素周期律，为以后发现新元素和波耳建立原子模型指明了方向。20 世纪 30 年代，积累了 100 多万种化合物的数据，结合量子化学的发展，导致鲍林提出共价、电价和氧化值的定义，以及 σ 键、π 键、杂化轨道、电负性、共振结构等概念，总结出化学键理论，发表《论化学键本质》这本经典著作，对 20 世纪化学的发展起了非常重要的作用。现在截至到 1999 年 12 月 31 日，美国《化学文摘》登记的分子、化合物和物相的数目已超过 2340 万种，比鲍林总结化学键理论时扩大了十余倍，但全世界的化学家似乎还没有充分利用这一化学文选宝库来总结规律。这是世纪之交的难得机遇，不可交臂失之"。

一般说来，数据库里的知识发现（Knowledge Discovery in Database，KDD），是指从大量的数据中提取出有效模式的非平凡过程，该模式是新颖的、可信的、有

效的、可能有用的和最终可以理解的[2]。而数据挖掘（Data Mining，DM）被认为是 KDD 中的一个步骤，是指利用某些特定的知识发现算法，在一定的运算效率限制下，从数据库中提取出感兴趣的模式[3]。数据挖掘技术无论在理论上，还是在实用技术上，都已取得了较大的进展[4~11]，同时也开发出了各种专用或通用的商业数据挖掘软件[12~16]。

化学化工领域积累了大量的科学实验和生产实际数据，如何总结这些数据中的规律性，进而用以指导以后的科学实验和生产操作，这是一项非常有意义的工作，这项工作的实施需要数据挖掘技术与化学化工知识和科学实践的结合。化学化工数据的不断积累是化学数据挖掘方法和技术应用的基础，而数据挖掘方法和技术的成功应用，一方面使我们更加认识到数据及其数据库的宝贵价值，促进数据采集和数据库技术的发展；另一方面对数据挖掘理论和算法不断提出新课题，促进计算机化学、化学计量学和化学信息学等新学科的发展。

化学数据挖掘方法和技术的应有领域非常广泛，下面结合我们在材料设计、构效关系和工业优化等方面的工作探讨化学数据挖掘的目的和意义。

1.1.1 数据挖掘与材料设计

新材料、新物质的探索和研制历来都是用经验方法，或称为“炒菜”（Trial and Error Method）式方法。即当要求提出后，凭经验决定材料制备的配方和工艺，制备一批样品，分析其成分和组织结构，测定其性能，若不合乎要求，则另行试制，一般要求反复多次才能获得成功。成功以后，还要摸索批量生产的技术和工艺条件，以实现廉价、批量生产的目的。这种“咸则加水，淡则加盐”的摸索方式虽然有效，但是终究事倍功半，费时费力。

为了摆脱这种较为盲目的研制方式，科学家们于 20 世纪提出了“材料设计”（Materials Design）的设想。所谓的“材料设计”，是指通过理论与计算预报新材料的组分、结构与性能，或者说，通过理论设计来“定做”具有特定性能的新材料。

1995 年，美国国家科学研究委员会（National Research Council，NRC）邀请众多专家进行调查分析，编写了《材料科学的计算与理论技术》这一专门报告，其中说：“Materials by Design”（设计材料）一词正在变为现实。日本学者在 1985 年提出了“材料设计学”一词。我国 1986 年开始实施“863 计划”时，对新材料领域提出了探索不同层次微观理论指导下的材料设计这一要求，从那时起，在“863 计划”材料领域便设立了“材料微观结构设计与性能预测”研究专题。所以，虽然用语有所差别，但关于材料设计的基本含义是共同的。

材料设计可按研究对象的空间尺度不同而划分为三个层次：微观设计层次，空间尺度约在1nm量级，是原子、电子层次的设计；连续模型层次，典型尺度约在1μm量级，这时材料被看成连续介质，不考虑其中单个原子、分子的行为；工程设计层次，尺度对应于宏观材料，涉及大块材料的加工和使用性能的设计研究。这三个层次的研究对象、方法和任务是不同的。

由于材料设计的研究对象多为由众多原子组成的复杂体系，原子间的作用复杂多样，难以用简单的解析方程求解，虽然原则上可以通过量子力学、统计力学的方程求解，但是仅从“第一原理”推算来把握复杂的材料设计体系和过程，在可以预见的未来尚难办到。与此同时，伴随着人类对新材料的开发和研制，积累了大量的数据，特别是近几十年来，随着信息技术的发展，各种有关材料性能和研制的数据库应运而生，互联网技术使得这些数据的获得也更为方便快捷。在这些海量的数据中隐藏着一条规律：何种原子或配方，按何种方式堆积或搭配，具有何种特定的物理和化学性质，即结构（或配方）-性质（性能）的关系。若能利用“第一原理”或者基础实验，根据已知的实验结果，找出目标值（性质或性能）与相关参数（原子参数、分子参数、工艺参数、成分含量等）的关系，总结出经验或半经验规律，并用于指导实验开发和提供材料设计的线索，即可以达到减少工作量，减少盲目性，解决实际问题的目的。

运用数据挖掘方法，对材料设计的相关数据加工处理，建立辅助新材料、新物质研制的专家系统，正在成为新材料设计的主流。有些专家系统已经用于新材料、新物质生产的优化控制，“材料智能加工系统”（Intelligent Processing of Materials）也在若干材料的研制和优化控制中试用成功。今天，材料设计不仅仅是科研院所的重点研究项目，也成为企业界的关注对象。计算机辅助材料设计大致可分为三个层次：

（1）材料、药物、染料、催化剂等的微观结构与性能的关系，从量子化学、固体物理、结构化学等角度探索研制新材料、新物质的新思想和新概念；

（2）从相图、热力学和动力学性质出发，探索新型合金、陶瓷等材料及其制备方法的革新；

（3）运用模式识别、人工神经网络、遗传算法、支持向量机等数据挖掘方法，结合数据库和知识库，总结材料结构与性能（性质）的关系、配方及工艺条件与材料性能或生产技术指标（成品率、能耗等）的关系等规律，用于材料制备和加工的优化。

利用化学数据挖掘方法和技术，可以总结新材料的物理化学性质与其组成元素的原子参数、化学配方、制备工艺等参数之间的定性或定量关系，在此基础上可以

辅助新材料研制和新产品开发，达到“事半功倍”的效果。

定量结构-性能关系研究（QSPR）为材料学的重要组成部分。研究者从材料的组成、结构特征和加工条件入手，利用数据挖掘方法可以总结和预测材料的具体性能。在实际应用中定量结构-性能关系的一些研究成果，可以指导材料的设计与生产流程，控制产品的合成路线，最终得到令人满意的结果。图 1.1 为现代材料设计与制备的基本流程。

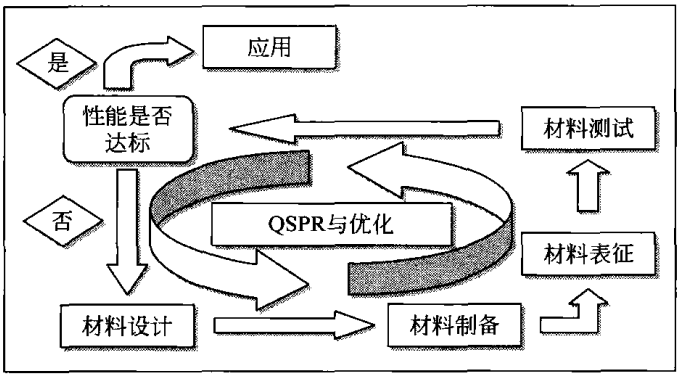


图 1.1 现代材料设计与制备的基本流程

1.1.2 数据挖掘与构效关系

化合物的性质/活性是化学的基本研究内容之一。化学家们普遍认为，化合物所表现出来的各种性质/活性与化合物的结构密不可分，即性质/活性是结构的函数。这也是结构-性质/活性关系（Structure Property/Activity Relationship，SPR/SAR）的基本假设，它们之间的关系如图 1.2 所示。结构-性质/活性关系也是化学的一个研究热点。1842 年，德国化学家 Koop 认为一系列相关化合物的物理化学性质可以由它们在一个矩阵中的位置得到预测，进而人们发现化合物拓扑结构是决定其化学性质的重要因素。1863 年，法国斯特拉斯堡大学的 A. F. A Cros 观察到，醇类物质对哺乳动物的毒性随着其水溶性的降低而增加。19 世纪 80 年代，德国马尔堡德大学的 Hans Horst Meyer 和苏黎世大学的 Charles Ernest Overton 分别独立地指出，有机物质的毒性与其亲油性相关。

20 世纪 40 年代起，化学家开始发现分子和其它化学物质可以很方便地用多种不同的矩阵表示[17,18]，化学图的概念及拓扑指数（图论指数）[19,20]的引入使表征分子结构并进行化合物的构效关系研究有了一个基本工具。

1964 年，Hansch 等从取代基与活性的关系出发，建立了线性自由能关系模型（Linear Free Energy Relationships，LFER），从而使定性研究定量化。Hansch 是该领域一系列概念和方法的提出者，对 QSPR/QSAR 研究的发展做出了重要贡献。

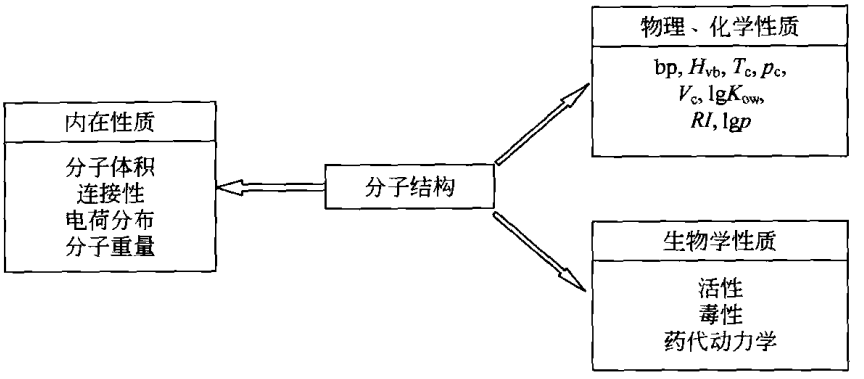


图 1.2 分子的构效关系

与此同时，Free 和 Wilson 提出了取代基贡献模型。近年来，随着化学计量学的发展，SPR/SAR 的研究提高到了一个新的水平。一方面，表征分子的结构参数不断丰富[21]；另一方面，一些新的建模方法也被引入到 SPR/SAR 的研究中[22,23]。结构-活性（性质）关系研究已成为化学、药物化学、环境化学中的一个前沿课题。

随着对分子结构的深入认识，以及适用的数理统计方法和计算机技术的引入，QSPR/QSAR 研究开始向三维发展，先后提出了 QSPR/QSAR 的位穴模型和比较分子场方法等，不仅取得了令人欣慰的成果，而且开辟了更为广阔的应用前景。

QSPR/QSAR 的研究同时渗透和推动了分子结构研究的深入，并取得了一系列重要突破，先后引出了很多新的结构参数。如拓扑结构指数中，除早期的 Wiener 径数（1947）和 Gordon-Scantlebury 指数（1964）等外，20 世纪 70 年代又提出了 Hosoya Z 指数（1971）、Balaban B 指数（1979）和 Kier 分子连接性指数 χ（1976）。80 年代 Simon、Crippen 等人又引入了一系列三维结构参数。这些参数大大丰富和促进了 QSPR/QSAR 的发展。因此，从分子结构出发，运用多元回归、人工神经网络、支持向量机等数据挖掘方法，可以总结分子的生物活性（或性质）与分子描述符（如物理化学参数、拓扑参数、几何参数以及量子化学参数）之间的关系，在此基础上可以设计和预测新的高活性化合物，然后再对预测的化合物进行化学合成，提高所需新化合物的命中率。

1.1.3 数据挖掘与工业优化

石油和化工企业是我国的基础支柱产业，在我国国民经济中占有举足轻重的地位。但与世界石化工业生产水平相比，我国的石油和化学工业还有不小的差距，例如我国乙烯生产的现金操作费用每吨约为 142 美元，比世界先进水平高出 24%，比亚太地区高出 5%。因此，如何利用工业优化技术提高劳动生产率和资源利用率，全面提升我国石油和化学工业的盈利能力和竞争能力，对于我国石油和化学工

业的可持续发展有着十分重要的意义。

提升企业的生产水平可以从设备改造、工艺改进等方面着手，实践证明虽然这些措施可以取得非常好的效果，但周期长、投资大。与此相比，利用控制技术和化学数据挖掘技术对生产操作进行优化，实施简便、见效快、投资回报率高，正越来越得到业界的重视。近年来，分布式控制系统（DCS）已经广泛应用于我国大中型石化装置，为试点和推广国内外新技术打下了基础。目前世界上已有 20 多家公司推出了 30 余种石化优化软件，应用领域遍及主要石化装置，其中先进控制（Advanced Process Control，APC）技术已经在我国几十个生产装置实施，如常减压、催化裂化、催化重整、加氢裂化、聚丙烯、聚乙烯等。根据 Chemshare 公司的调查结果，在已有 DCS 系统基础上实施先进控制的投资收益比为 1∶4，在先进控制基础上实现装置实时优化的投资收益比也为 1∶4。因此，先进控制和实时优化控制挖潜增效的效果非常明显。

为了从生产机理上建立描述过程的精确模型，以谋求更好的优化效果，基于机理模型的石化优化软件应运而生。这类软件主要用于过程模拟、装置设计及实时优化控制。过程模拟软件通常利用物理化学原理进行工艺计算、物性计算、能量和质量平衡计算等，软件中采用了回归分析、数据拟合等数理统计方法。机理模型通常有较高的精度，可以在计算机上模拟实际生产装置的某些特性，是设计人员在生产装置没有建立之前预测或验证设计的重要工具。

近年来，基于数据挖掘的工业优化技术已在国外受到高度重视，应用的案例日益增多。数据挖掘技术用于生产优化可与先进控制、实时优化控制互为补充，相得益彰。化工生产过程涉及许多复杂的物理、化学变化，常常很难通过机理来建立模型，即便建立了模型，其精度也很低，模型只能用来表明生产的大体变化趋势，而无法用来指导生产。此外，工业生产过程中存在许多可变因素和干扰（原料性质、设备状态、操作工况的变化，生产环境和生产系统自身的干扰），数学模型通常是在某一特定条件下建立的，因而仅仅在小范围内适用，在实际复杂多变的生产中难以使用。随着计算机科学和过程系统工程的发展，工业生产过程自动化程度越来越高，工业生产数据采集和存储越来越经济便利，对于一个中等规模的石化生产装置，其 DCS 系统的仪表位号点数约 500 点，如果每分钟保存一个生产数据，那么，每天就有 70 万个生产数据，一年可达 2.5 亿个数据。这些数据记录了工业生产过程的特征、性能、变化等，是生产装置的本质反映。利用数据挖掘技术，可以从工业生产数据中寻找规律和发现知识，并用这些知识指导企业的生产过程，从而达到优化生产过程，使企业效益最大化。

传统上，研究者用统计图表总结生产数据，但这种统计图表不能提供有关生产