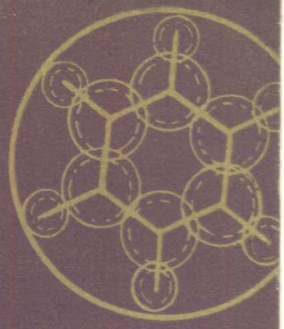
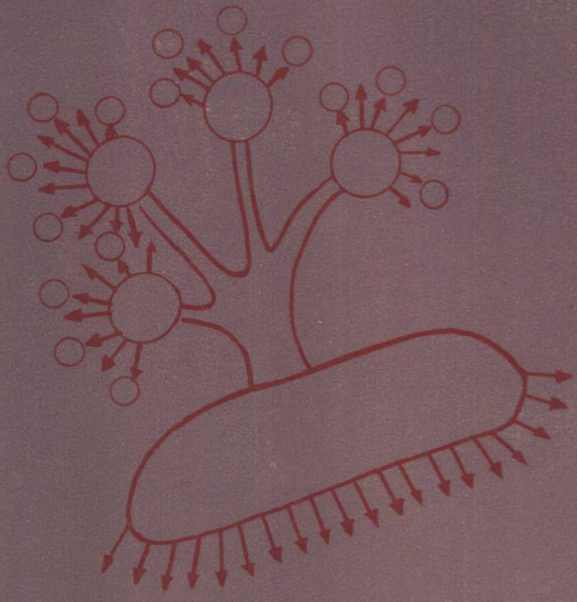
JDHXCS

近代化学丛书

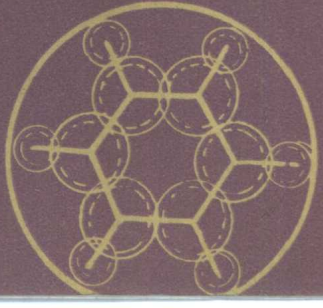




计算化学及其应用

陈念贻　许志宏　刘洪霖　徐　桦　王乐珊　著

上海科学技术出版社



近代化学丛书

计算化学及其应用

陈念贻　许志宏　刘洪霖
徐　桦　王乐珊　著

上海科学技术出版社

近代化学丛书

计算化学及其应用

陈念贻　许志宏　刘洪霖
徐　桦　王乐珊　编著

责任编辑　周贞瑛

上海科学技术出版社出版
（上海瑞金二路 450 号）

新华书店上海发行所发行　江苏溧水印刷厂印刷

开本 850×1156 1/32　印张 15.125 字数 399,000
1987 年 10 月第 1 版　1987 年 10 月第 1 次印刷
印数：1—3,600

书号：13119·1376　定价：3.05 元

《近代化学丛书》编辑委员会

序

化学是自然科学的主要基础学科之一．因此，化学科学在国民经济和科学技术领域中占有极其重要的地位，是农业、能源工业、材料科学、计算机工业、激光技术、空间技术、高能物理和遗传工程等不可缺少的基础．由于学科之间的相互渗透和交叉，以及新的实验手段和计算机的广泛应用，大大推动了近代化学的发展，产生了许多分支学科和边缘学科，如计算机化学及其应用、激光化学、量子有机化学等等，所以，化学科学正处在一个崭新的发展时代．

建设社会主义四个现代化和化学学科赶超世界先进技术水平，需要千千万万个有才干的化学工作者为之共同努力．为了更好地进行工作，并取得成果，必须具有渊博的知识，了解化学科学的发展现状和动态，善于吸取相邻学科的新成就，牢固地掌握近代化学的基础理论．因此，对青年化学工作者的造就和培养，就显得十分紧迫和重要．为此目的，我和卢嘉锡等十五位同志应上海科学技术出版社的要求，组织编写了《近代化学丛书》．

该《丛书》按专题比较系统地、深入地论述某一领域的基础理论，是一套具有较高理论水平的著作．

该《丛书》注意理论联系实际，在论述基础理论的同时，注重结合教学和科研工作，反映近代化学发展的最新成果，以供高等院校有关专业高年级学生、研究生、教师及有关科研和工程技术人员参考．

该《丛书》包括近代无机化学、理论有机化学、量子有机化学、稀土物理化学、界面及胶体化学、原子簇化合物、计算机化学及其应用、表面化学、半导体物理化学、金属有机在有机合成中的应用、

有机催化、激光化学、海洋化学、地球化学等内容，分册陆续出版。

还应说明，该《丛书》的著作者，虽有良好的写作愿望和积极性，并在教学和科研方面具有丰富经验，但由于阅历和理论水平不同，各分册之内容深浅、繁简取舍等很难取得统一。同时，还有可能存在一些不妥，甚至错误之处，敬希读者谅解并予以指正。倘若该《丛书》能成为广大化学工作者确有参考价值的基础理论读物，对四个现代化建设有所贡献的话，那么，我们组织编写这套《丛书》的目的就算达到了。

唐敖庆

一九八二年十二月于长春

前　言

现代化学发展的主要趋势之一，是计算机在化学各个分枝的广泛应用．计算化学即是由此产生的一门新学科．由于计算机在化学中应用的广泛性和多样性，也由于这方面近年来发展十分迅猛，很难对计算化学的学科范围作一个明确的界线，也很难写一本包罗万象的计算化学著作．

本书各章的作者都是多年从事计算化学的研究人员．在研究工作中积累了丰富经验，也收集了大批国内外文献资料．本书各章就是在这个基础上写成的．因此这本书不是一本面面具到的计算化学教科书，而是对计算化学四个重要领域——模式识别在化学中的应用、计算机模拟在化学中的应用、量子化学计算方法和化学数据库——作较深入地介绍的一本学术著作．本书部分内容对应用化学研究和化工生产有用．另外一部分内容则对化学的基础理论研究有用．因此本书对化学领域的理论研究人员和实验工作人员，以及应用化学工作者都有参考价值．

在本书各章有关的研究工作中，我们得到徐光宪教授、唐敖庆教授和唐有祺教授的指导和帮助．江乃雄、谢雷鸣、缪强、李通化、张未名等同志在工作中做出许多成果，丰富了本书的内容．我们在此一并表示感谢．

希望广大读者对本书内容提出宝贵意见，以便在今后补充或改正．

陈念贻

1986 年于上海

内 容 简 介

本书介绍了计算化学的四个重要领域：化学模式识别，化学体系的 Monte Carlo 法和分子动力学法计算机模拟，计算量子化学和化学数据库．各章作者结合自己的科研成果，介绍了这四个重要领域的科学原理、计算方法和应用实例．

本书对象是化学、计算机应用以及化工、冶金等有关学科的研究人员、大学教师、工程技术人员、研究生与高年级大学生．

目　录

第一章　化学模式识别 ……1
§1.1　模式识别的基本概念……1
§1.2　模式识别的计算方法 ……12
§1.3　模式识别与金属材料研究 ……34
§1.4　模式识别在氧化物系研究中的应用 ……56
§1.5　模式识别与催化研究 ……89
§1.6　模式识别与谱图解析……101
§1.7　模式识别的其他应用……112
参考文献 ……131
第二章　Monte Carlo 方法及其在化学中的应用 ……137
§2.1　Monte Carlo 方法的基本原理 ……137
§2.2　古典液体和溶液的 Monte Carlo 法的计算机模拟——I. 硬球模型模拟……144
§2.3　古典液体和溶液的 Monte Carlo 法计算机模拟——II. 离子液体模拟 ……150
§2.4　古典液体和溶液的 Monte Carlo 法计算机模拟——III. 原子及分子液体……175
§2.5　Monte Carlo 法在高分子科学中的应用……190
§2.6　Monte Carlo 法在其他化学问题中的应用……198
参考文献 ……206
第三章　分子动力学方法 ……209
§3.1　分子动力学方法的原理……209
§3.2　分子动力学方法的应用……244
参考文献 ……271
第四章　计算量子化学 ……274
§4.1　计算量子化学的基本原理……274

§4.2 从头计算方法……292
§4.3 简化的从头计算方法……321
§4.4 半从头计算与半经验方法……337
§4.5 超 Hartree-Fock 计算方法……352
§4.6 应用……363
参考文献……386
第五章 化学数据库……392
§5.1 化学数据库的产生……392
§5.2 几种典型化学数据库……405
§5.3 开发化学数据库的基本过程……445
参考文献……466
附录一 方位阱势在 σ_2 碰撞的动力学……469
附录二 Ewald 方法……470

第一章　化学模式识别

陈念贻

§1.1　模式识别的基本概念

1. 计算机模式识别方法——问题的提出

在科学研究或生产活动中，人们经常需要对某些未能直接观测(或尚未直接观测)的事物作出估计或预报. 例如：化学家需要估计一个未知样品的分子结构或化学组成，材料科学家需要估计一批尚未制备的新材料的物理、化学性质，药物学家需要估计某一类化合物的结构和疗效的关系等等. 如果人们对被估计或预报的事物的理论知之甚详，也许可以对未知事物作严格的理论预测. 可惜在大多数情况下，或是由于研究对象过于复杂，或是由于其详细机理不十分清楚，或是由于理论计算工作量太大以致难于实施，严格的理论预测难以实现. 这就需要从大量已知实验事实中总结经验规律，进而估计或预报未知. 而某种程度的理论知识则是总结经验规律的重要依据.

在许多情况下，由于被估计或预报的事物不是由一种因素决定的，而是与多种因子有错综复杂的关系，单靠人的思维往往难于有效地抽提经验规律. 这时，计算机信息处理技术往往能帮助我们总结规律，预报未知，解决问题. 计算机模式识别(computerized pattern recognition)技术便是这样一种有用的手段[1~4].

试举一例：在鉴定一个未知化合物时，人们往往对它做各种谱学测量，例如做红外光谱、极谱、色谱、波谱等等，并根据这些谱综合判断未知化合物的结构式. 由于化合物及其谱学数据极其众多，做这种综合判断即使是有经验的专家往往也非易事. 但是现

在有了计算机模式识别方法，和数据库相结合，已能自动化地鉴定未知化合物样品的结构了.

再举一例：人们已积累了大批的物质的化学结构数据和物理、化学特性或生物活性的数据. 却常常未能详尽地查明结构与性质关系的详细机理. 为了掌握结构与性质关系的规律，以便更有效地合成、试制新材料、新药物，可用计算机模式识别方法总结经验规律，预报未知物质特性，更有目的地研制新物质.

用计算机模式识别方法总结规律预报未知，虽然不如彻底的理论方法那样严格和漂亮动人，却远比彻底的理论方法更富有适应性，因而能够在现有知识不甚完整的情况下发挥作用. 计算机模式识别方法和彻底的理论方法是相辅相成的，人们一般需要有关于研究对象某种程度的理论知识，可以估计(或至少可以猜测)影响研究对象的大致有哪些因素，用这些因素的参数构筑多维空间，才能用模式识别方法总结规律. 另一方面，由模式识别方法找出的规律只是经验规律，只是"知其然不知其所以然". 但这往往能给理论研究工作者以某种启示，通过理论研究找出真正的机理. 计算机模式识别方法离不开大量实验数据. 计算机本身不能产生信息，它只是信息的"加工厂". 人们根据已知实验事实的集合(称为"训练点")来"训练"计算机，使其"认识"规律，才能预报未知. 然而由于计算机往往能更有效地从训练点中抽提信息，所以能指导我们下一步的实验少走弯路，收到事半功倍之效. 所以运用模式识别方法常可节约大量的实验工作和人力、物力. 而计算机的预报也常不能百分之百准确，还需要实验检验，丰富了训练点，甚至改进预报用的程序. 如此，实践、认识、再实践、再认识……循环往复，就会使我们的知识愈来愈完全，预报方法愈来愈准确. 因此，用计算机模式识别技术帮助我们认识客观事物，是一种"归纳""演绎"相结合的方法，是完全合乎辩证唯物论认识论的原理的.

2. 统计模式识别和句法模式识别

计算机模式识别方法大致可分为统计模式识别 (statistical

pattern recognition) 和句法模式识别 (syntactic pattern recognition) 两大类. 统计模式识别将每个样本(或模式)用特征参数表示为特征量空间中的一个点，根据各点间的距离或距离的函数来判别、分类，并利用分类结果预报未知. 统计模式识别发展较早，在化学上已有了广泛的应用. 这构成了本书有关模式识别章节的大部份内容.

句法模式识别则是近十年来迅速发展的一类新的模式识别方法. 和统计模式识别不同，句法模式识别着眼于突出模式结构的信息，采用形式语言理论来分析和描述模式的结构. 这类方法已在遥感图片处理、指纹分析、汉字识别等方面广泛使用. 但在化学上应用的实例尚少. 由于句法模式识别更便于处理图形和结构信息，可以想见它在化学领域很有应用的潜力，本书对此也将作些讨论.

3. 模式空间(pattern space)

统计模式识别方法可用直观的参数空间讲述. 这种讲法易为化学工作者接受，故本书采用这一讲述方法.

可从一简单实例[5,6]说起：图 1-1 是一张化学键参数图，用以描述二元系的金属键化合物和共价-离子键化合物的分野. 纵坐标是二元系的元素的电荷-半径比(价电子数 Z 和原子实有效半径 r_k 之比)和 $\sum Z/r_k$，横坐标是二元系的电负性差. 可看出金属键化合物(以不服从原子价律为标志)和共价-离子键化合物分布在不同区域，其间有明显的分界线. 从图 1-1 还可看到一个本征半导体的分布区. 用图 1-1 可对二元系中化合物的化学键类型作估计.

图 1-1 实际上就是统计模式识别方法的某种雏型：在二维的图中，每一二元系代表点由 $\sum Z/r_k$、Δx 两个参数规定. 该二元系就是一个"样本" (sample)，坐标参量 $\sum Z/r_k$、Δx 称为特征量 (features)，样本表征为二维空间 $\Phi(\sum Z/r_k, \Delta x)$ 的一个点. 此二维空间即是此处的模式空间 (pattern space)，每个样本的参量集

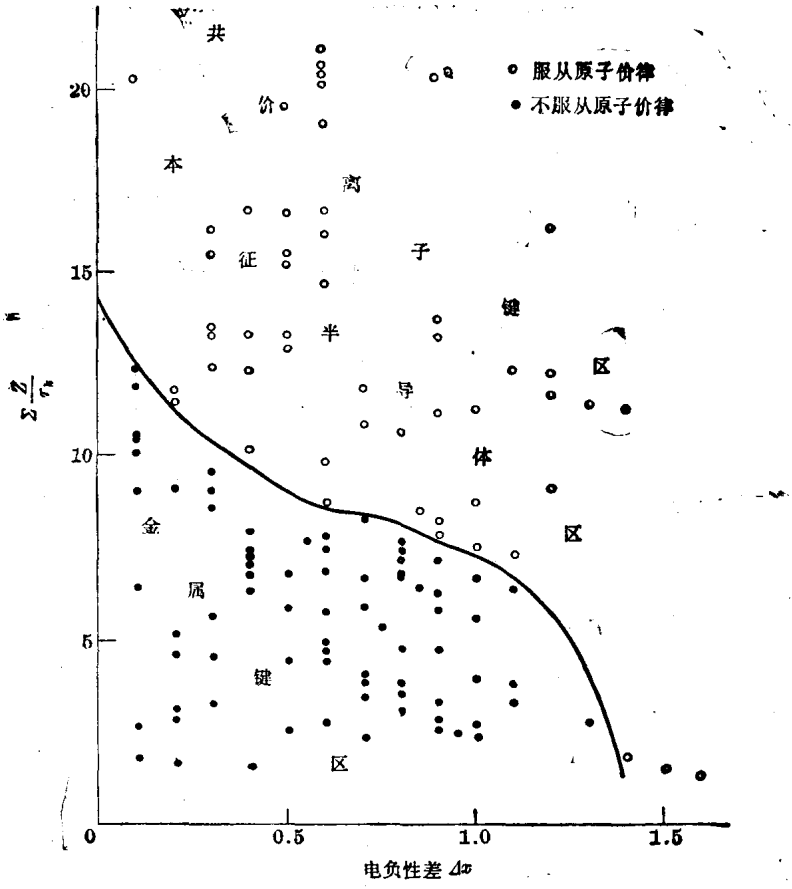


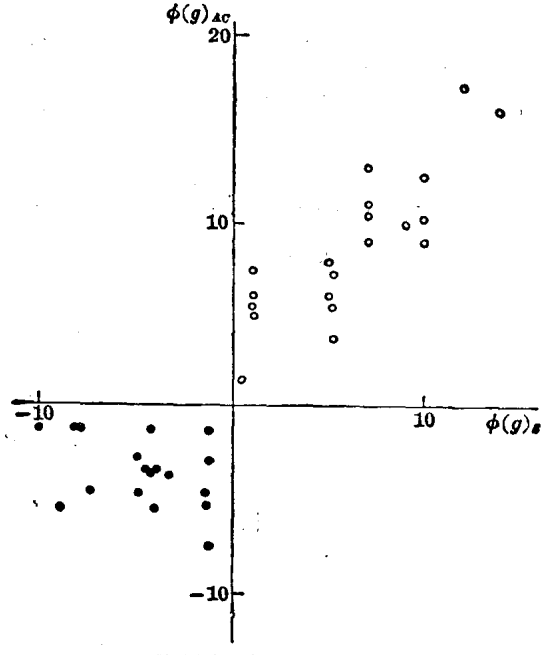
图 1-1　金属键的形成条件

合构成一个模式(pattern)，代表点的向量表示是模式向量(pattern vector)。

由于这个课题碰巧可用二维空间即平面图方法解决，不一定要用计算机信息处理。但实际上化学中的大多数问题，其决定性因素往往不止两个，用二维作图法并不是总能奏效。这类问题一般而言牵涉到高维的模式空间。多维空间中点列分布的图象的识别是个复杂问题。因为人类对高维空间缺乏感性认识，所以最好用计算机模式识别方法解决问题。举例说，图 1-1 处理的是二元

系，如果用类似参数组处理三元系信息，就要用三个 $\sum Z/r_k$ 和三个 Δx 张成的六维空间．碰巧有时可找到能区分两类或多类样本的多维空间的剖面，将多维空间的问题还原为二维空间处理．例如我们曾用三元系的两个 $\phi(g)$ 函数作纵横坐标作图，部份解决了三元化合物键型问题(见图 1-2)．但一般而言，这不是解决问题的好方法．因为 ① 许多多维空间的点列未必能通过简单地投影而有效地分类，② 即使客观上存在能有效分类的二维剖面，要用人工找出这种剖面也是很麻烦的事情．所以，在大多数化学课题中，模式识别都要处理多维空间的信息．多维空间信息处理的原理和二维空间信息处理并无根本区别，只是由于人类对多维空间缺乏直观想象能力，所以需要用抽象的数学方法实现模式识别．

所有的统计模式识别都基于这样一条原则：有相似性质的样本在模式空间中互相接近并形成集团——即"物以类聚"的原则．



● 不服从原子价律　○ 服从原子价律

图 1-2　三元化合物的键型和 $\phi(g)$ 的关系

4. 二类分类与多类分类

模式识别中最简单的分类问题是互不包容的二类分类(即第一类包括具有某种特性的样本,第二类则包括所有其它样本). 若模式空间中能找到一个或一组超平面或超曲面加以区分, 则寻求分类判据的计算称为"训练"(training). 若模式空间中两类样本的代表点可用一个超平面加以区分, 则这些样本称为"线性可分"的. 该超平面称为判别平面.

判别平面通常用与此平面正交的判别向量 $\boldsymbol{W}$ 表示. 判别向量便于判别某一未知样本的代表点在平面哪一边. 无向量乘积 $S=\boldsymbol{W}\cdot\boldsymbol{X}$($\boldsymbol{X}$ 是模式向量)为正值时代表点属于第一类,为负值则属于第二类. 对于多维空间,较方便的是用下列方程式

$$S=\sum_{i=1}^{d} W_i X_i$$

式中 S 为无向量乘积, X_i 为模式向量 $\boldsymbol{X}$ 的分量(亦即代表点的坐标值,或样本的一个表征参数值), W_i 为判别向量 $\boldsymbol{W}$ 的分量, d 为模式空间维数(亦即每个样本的表征参数的数目).

在线性可分的情况下,判别向量的计算需要用电子计算机. 一旦判别向量为已知, 具体课题中未知样本的分类(即预报)就只用手算就行了. 许多线性不可分的多维空间点列可通过适当的空间变换化为线性可分. 一种办法是不但取 X_i 等 $\boldsymbol{X}$ 的分量为坐标,而且取种种含 X_i、X_j 的高次项例如 X_i^2、X_iX_j、X_j^2 等为坐标. 用这种人为地增加空间维数的方法, 原则上可使多维的超曲面变换为超平面. 这样就可找出适当的判别平面. 当然, 当样本点的分类情况过于复杂时,上述算法非常麻烦,以致于实际上不行.

需要将样本区分为多类的分类(多类分类)在化学上也是常见的情况(例如: 同一价型的化合物往往有多种晶型,在这种情况下,晶型的分类就是多类分类). 多类分类远较二类分类复杂. 但有时可用二类分类法处理多次. 先将整套点列分成两族, 再从两族中进一步分类. 重复适当次数即完成多类分类. 多类分类也可采

用其它专门的方法. 例如 K 最近邻法(KNN 法)就能直接做多类分类.

5. 分类方法的试验和评估

模式识别方法的种类很多,常用的例如非线性映照法,线性判别函数法、K 最近邻法、SIMCA 法等等. 常常有人问:哪种方法分类效果更好?笼统地说不可能有统一的答案——不同的情况适用不同的方法. 在实用过程中,往往通过实际数据的分类效果来评估分类法的优劣.

在有一大批已知样本数据的基础上进行模式识别分类研究时,通常先将已知样本的数据随意分为二部份,一部份的样本(或其代表点)称为训练点集(training set),用模式识别方法对训练点集进行分类. 此时的正确分类率称为识别率(recognition rete). 如识别率够高,即可认为已建立了判别方法(例如:判别向量等等). 用这一判别方法对其余的样本(它们在"训练"中未"告诉"计算机,故对于计算机可称为"未知"样本)分类. 此时的正确分类率称为预报能力(predictive ability).

当然,在实用过程中,也可少留甚至不留未知样本,以增加训练点集,提高预报能力. 一种常用的办法是"留一法"(leave one out method),即每次取去一个样本,以其余样本的代表点作训练点,并将求得的预报方法对取去的一个样本的类别作"未知"预报. 这样依次对每个样本都作了"未知"预报后,取预报成功率(平均值)作为平均预报能力. 与此相似,也可有"留十法""留四分之一样本法". 用所有已知样本作训练点,用实验方法验证预报的可靠程度,当然是更为令人兴奋和信服的方法.

6. 有人管理分类 (supervised classification) **和无人管理分类** (unsupervised classification)

科学上的分类问题有两种. 一种是预先规定好分类的标准和种类的数目,通过大批已知样本的信息处理(即"训练")找出规律,