第 1 章　绪　论

1.1　数据挖掘技术简介

1.1.1　数据挖掘的背景介绍

随着数据库、数据仓库技术的发展和 Internet 的迅速普及，人们积累的数据量以前所未有的速度急剧增长，无论商业、企业、科研机构还是政府部门都积累了大量的、以不同形式存储的数据资料。在天文学研究、粒子物理学研究、化学研究、医学研究以及政府统计等方面出现了很多大型数据库。最近，人类基因组计划的实施和后基因组计划的到来使得数据更加丰富多彩。然而，在拥有海量数据的同时，我们对数据中所蕴涵的信息和知识缺乏充分的理解和应用，出现了“数据丰富而信息缺乏”的严峻局面，依靠传统的数据库对数据进行查询和检索等分析手段很难帮助用户从数据中提取带有指导性和结论性的有用信息。虽然基于数据仓库的联机分析处理（On-Line Analysis Process，OLAP）技术具有总结、概括和聚集的功能，支持多维分析和决策，但它不能进行更深层次的数据分析，数据库中蕴藏的丰富知识得不到充分的挖掘和利用。因此，人们迫切需要新的强有力的数据分析方法和技术以解决“数据丰富而信息缺乏”这一现象，帮助人们从繁杂的数据中挖掘出有用的信息，发现其中存在的关系和规则，实现决策的智能化和自动化，从而带来商业上巨大的信息价值。在这种情况下，数据挖掘（Data Mining，DM）技术应运而生，并显示出强大的生命力。

1989 年 8 月，在美国底特律市召开的第十一界国际人工智能联合学术会议上正式形成“数据挖掘”一词，它常常与数据库中的知识发现（Knowledge Discovery in Database，KDD）一起使用。从 1995 年开始，每年举办一次 KDD 国际学术会

议，将 KDD 和 DM 方面的研究推向了高潮，从此，“数据挖掘”一词开始流行。在中文文献中，DM 有时还被翻译为“数据采掘”、“数据开采”、“数据发掘”等。

1.1.2 数据挖掘的研究现状

1989 年 8 月，在美国底特律市召开的第 11 届国际联合人工智能学术会议（UCAI-89）上，Gregory Piatetsky-Shapiro 组织了“Knowledge Discovery in Database”专题讨论会，该讨论会的重点是强调发现（Discovery）的方法以及发现的知识（Knowledge）两个方面，这是基于数据挖掘概念的首次国际学术会议。

Gregory 在为该会命名时，曾考虑过数据挖掘（Data Mining）、知识挖掘（Knowledge Mining）、知识抽取（Knowledge Extraetion）和数据库挖掘（Database Mining）这四个术语。但是，因为数据挖掘已经在数据库领域被使用过，所以不具备吸引力，而且统计学界对“数据挖掘”这个术语抱有偏见，认为挖掘的语义太单调，并且没有指明挖掘的内容。“知识挖掘”和“知识抽取”并不比“数据挖掘”更理想，而“数据库挖掘”是 HNC 公司的注册商标（Database Mining TM），因此，最终选择了 KDD 作为专题讨论会的名称。随着该学术会议的召开，KDD 开始在人工智能（Artificial Intelligence，AI）和机器学习（Machine Learning，ML）领域变得流行。

学术界有部分观点将 KDD 作为知识发现的整个过程的统称，而将数据挖掘作为 KDD 过程中的一个步骤。然而数据库领域的研究人员经常与商业伙伴接触，非学术人员（商业伙伴）更容易接受“数据挖掘”这一术语，因此，数据挖掘在商业出版社比 KDD 或其他名词更为流行。目前，数据挖掘的研究已经不仅仅局限于数据库中的数据挖掘，挖掘的对象包括文本、图像、声音等多种非结构化数据格式，从这个意义上讲，数据挖掘比 KDD 更贴近这一术语本身代表的含义。到目前为止，数据挖掘和 KDD 这两个术语几乎是同义的，本文所指的数据挖掘也是取其广义的概念，即表示知识发现的整个过程，而不仅仅是应用算法这个步骤。

随后在 1991 年、1993 年和 1994 年都举行了 KDD 专题讨论会，来自各个领域的研究人员和应用开发者集中讨论了数据统计、海量数据分析算法、知识表示

和知识运用等问题。随着参与科研和开发人员的不断增加，国际 KDD 组委会于 1995 年把专题讨论会发展成为国际年会。在加拿大的蒙特利尔市召开了第一届 KDD 国际学术会议，在这次会议上，“数据挖掘”（Data Mining）概念第一次由 Usama Fayaad 提出。以后每年召开一次，其会议全称为 ACM SIGKDD （Special Interested Group on Knowledge Discovery in Databases）International Conference on Knowledge Discovery and Data Mining。参加人数由几十人发展到上千人，研究重点也逐渐从发现方法转向系统应用，并且注重多种发现策略和技术的集成，以及多种学科之间的相互渗透。1998 年，在美国纽约举行的第四届知识发现与数据挖掘国际学术会议上，与会者不仅进行了学术讨论，而且领略了 30 多家软件公司展示的数据挖掘软件产品。

除了美国人工智能协会主办的 KDD 年会外，还有许多的数据挖掘年会，包括 PAKDD、PKDD、SIAM-Data Mining 等。PAKDD（Pacific-Asia Conference on Knowledge Discovery and Data Mining）是亚太平洋地区数据挖掘年会，从 1997 年开始，每年召开一次，至今已召开了九届，其中，1999 年的 PAKDD 在我国北京召开，2007 年的 PAKDD 于 5 月 22～25 日在南京举行。PKDD（European Symposium on Principles of Data Mining and Knowledge Discovery）是欧洲数据挖掘会议，也是从 1997 年开始每年召开一次。SIAM-Data Mining（Society for Industrial and Applied Mathematics）是 SIAM 组织召开的数据挖掘讨论会，于 2001 年 4 月召开第一届，专注于科学数据的数据挖掘，以后每年召开一次，第七届 SIAM 数据挖掘国际会议于 2007 年 4 月 26～28 日在美国明尼苏达州的明尼阿波利斯市召开。

国外已经有许多专门的工作组从事数据挖掘领域的研究，比较著名的有 R.Agrawal 领导的 IBM Almaden 实验室的数据挖掘工作组；J • Han 带领的 SFU 工作组；Stanford 大学的 Ullman 领导的关联规则研究小组；Minnesota 大学的 Kumar 领导的并行数据挖掘研究小组；新西兰怀卡托大学 Ian H • Witten 教授领导的 Weka 工作组等。他们提出了许多好的数据挖掘算法，并实现了数据挖掘工具，为该领域的发展奠定了一定的基础。其中，Ian H • Witten 教授在 2004 年荣获了国际信息处理联合会（IFIP）颁发的 Namur 奖项，这是一个两年一届、用于奖励那些在信息和通信技术的社会应用方面做出杰出贡献及具有国际影响的

荣誉奖项。2005 年 8 月，在第十一届 ACM SIGKDD 国际学术会议上，Weka 工作组荣获了数据挖掘和知识探索的最高服务奖，Weka 被誉为数据挖掘和机器学习历史上的里程碑。

国内对 DM 的研究稍晚，没有形成整体力量，但最近几年有较大的发展。北京大学智能科学系的唐世渭和杨冬青教授领导开发了基于空间数据挖掘的客户分析系统模型 CASDM；复旦大学施伯乐教授领导开发了数据挖掘工具集 AMINER；清华大学石纯一和陆玉昌教授、中科院计算所史忠值研究员、北京科技大学杨炳儒教授、武汉大学李德仁院士和复旦大学周傲英教授等对数据挖掘算法进行了深入的研究；清华大学周立柱教授、复旦大学朱扬勇教授以及四川大学唐常杰教授领导的数据挖掘工作组取得了许多重要的研究成果。

1.1.3 数据挖掘的相关知识

1. DM 定义

DM 就是从大量的、不完全的、有噪声的、模糊的、随机的数据库中提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。这个定义包括以下四层含义：

（1）数据源必须是真实的、大量的、含噪声的。

（2）发现的是用户感兴趣的知识。

（3）发现的知识要可接受、可理解、可运用，最好能用自然语言表达发现结果。

（4）并不是要求发现放之四海而皆准的知识，也不是要去发现崭新的自然科学定理和纯数学公式，更不是什么机器定理证明，所有发现的知识都是对的，是有特定前提和约束条件、面向特定领域的。它是一门涉及面很广的交叉学科，包括机器学习、数理统计、人工智能、神经网络、数据库、模式识别、粗糙集和模糊数学等相关技术。

2. DM 的研究内容和任务

随着 DM 研究逐步走向深入，DM 和 KDD 的研究已经形成了三个强大的技术支柱：数据库、人工智能和数理统计。因此，机器学习、模式识别、人工智能领域的常规技术，如分类、聚类、决策树、神经网络、进化计算、模糊集和粗糙

集等方法经过改进大都可以应用于 DM。但是，数据挖掘系统通常面对的是大量的类型更加复杂的数据，因而，对现有技术的改进、综合各种方法技术优点的有效集成以及研究面向 DM 的新技术都是 DM 的研究内容。概括起来主要有基础理论（包括数据库、数据仓库以及海量数据的存储和调用）、发现算法（包括概括、分类、聚类、关联等针对特定挖掘任务和知识的有效方法）、知识表示方法和可视化技术、发现知识的维护和再利用、半结构和非结构化数据中的知识发现以及网络 DM 等。

按照 DM 技术所能发现的模式，可以将挖掘任务分为两大类：预测型（Predictive）和描述型（Descriptive）。

（1）预测型模式。根据数据项值精确确定某种结果的模式，挖掘预测型模式所使用的数据也都是可以明确知道结果的。

（2）描述型模式。对数据中存在的规则做一种描述，或者根据数据的相似性把数据分组。按照 DM 技术所能发现的规则，挖掘任务分成五类：①总结规则挖掘。从指定的数据中，从不同的角度或层次上挖掘出平均值、极小值、极大值、总和和百分比等；②关联规则挖掘。从数据库中挖掘出满足一定条件的依赖性关系；③分类规则挖掘。在已知训练集的特征和分类结果的基础上，为每一种类别找到一个合理的描述或模型；④聚类规则挖掘。客观地按被处理对象的特征分类，将有相同特征的对象归为一类；⑤预测及趋势性规则挖掘。对数据进行分类或回归分析，或对数据将来的发展趋势进行估计。

通过发现模式或规则，我们一般可以得到以下几类知识：①广义知识（Generalization）；②关联知识（Association）；③分类知识 （Classification and Clustering）；④预测型知识（Prediction）；⑤偏差型知识 （Deviation）。

3. DM 的过程

作为一个学术领域，DM 和 KDD 具有很大的重合度，大部分学者认为 DM 和 KDD 是等价的概念。相对来讲，DM 主要流行于统计、数据分析和数据库领域，而 KDD 则主要流行于人工智能和机器学习领域。从数据处理的过程看，可以把 DM 看作 KDD 过程中同算法相关的一步，借助于算法，在可接受的计算范围内从数据中枚举模式或模型结构。KDD 的基本过程包括数据准备（data preparation）、

DM、知识的解释和评估（interpretation and evaluation）。

（1）数据准备。包括确定要挖掘的数据对象，即数据源。搜索所有与业务对象有关的内部和外部数据信息，同时进行数据清理和整合。从数据集中选择出与分析任务相关的数据，研究数据的质量，为进一步的分析作准备，将数据转换成一个分析模型，这个分析模型是针对 DM 算法建立的，建立一个真正适合挖掘算法的分析模型是 DM 成功的关键。

（2）DM 阶段。对所得到的经过转换的数据使用智能方法进行挖掘，并从中提取出模式。

（3）知识的解释和评估。根据某种兴趣度度量，识别表示知识的真正有趣的模式，同时使用可视化和知识表示技术，向用户提供挖掘的知识。DM 是一个循环的过程，可能需要循环多次才有可能达到预期的效果。

4. DM 的主要方法

（1）神经网络方法（Artificial Neural Network，ANN）。

神经网络由于本身良好的鲁棒性、自组织自适应性、并行处理、分布存储和高度容错等特性非常适合解决 DM 问题，因此，近年来越来越受到人们的关注。典型的神经网络模型主要分三大类：以感知机、BP 反向传播模型、函数型网络为代表的，用于分类、预测和模式识别的前馈式神经网络模型；以 Hopfield 的离散模型和连续模型为代表的，分别用于联想记忆和优化计算的反馈式神经网络模型；以 ART 模型、Koholon 模型为代表的，用于聚类的自组织映射方法等。

（2）遗传算法（Genetic Algorithm，GA）。

GA 是一种基于生物自然选择与遗传机理的随机搜索算法，是一种仿生全局优化方法。它利用生物遗传学的观点，通过自然选择、遗传、变异等作用机制，实现各个个体的适应性的提高。这一点体现了自然界中“物竞天择、适者生存”的进化过程。1962 年，Holland 教授首次提出了“GA 算法”的思想，从而吸引了大批的研究者，迅速推广到优化、搜索和机器学习等方面，并奠定了坚实的理论基础。 用遗传算法解决问题时，首先要对解决问题的模型结构和参数进行编码，一般用字符串表示，这个过程也是把问题符号化、离散化的过程，也有在连续空间定义的 GA。GA 具有的隐含并行性、易于和其他模型结合等性质使得它在 DM

中应用广泛。

（3）决策树方法（Decision Tree，DT）。

决策树是一种常用于预测模型的算法，它通过将大量数据有目的地分类，从中找到一些有价值的、潜在的信息。它的主要优点是描述简单，分类速度快，特别适合大规模的数据处理。最有影响和最早的决策树方法是由 Quinlan 提出的著名的基于信息熵的 ID3 算法。

（4）粗糙集方法（Rough Set，RS）。

粗糙集理论是一种研究不精确、不确定知识的数学工具。粗糙集方法有几个优点：不需要给出额外信息；简化输入信息的表达空间；算法简单，易于操作。粗糙集处理的对象是类似二维关系表的信息表。目前成熟的关系数据库管理系统和新发展起来的数据仓库管理系统为粗糙集的 DM 奠定了坚实的基础。但粗糙集的数学基础是集合论，难以直接处理连续的属性，而现实信息表中的连续属性是普遍存在的。因此，连续属性的离散化以及直接使用连续属性数据应用粗糙集理论是一个难点。

（5）模糊集方法（Fuzzy Set，FS）。

模糊集方法，即利用模糊集合理论对实际问题进行模糊评判、模糊决策和模糊模式识别，聚类分析系统的复杂性越高，模糊性越强。一般模糊集合理论是用隶属度来刻画模糊事物的。

（6）支持向量机方法（Support Vector Machines，SVM）。

传统统计学理论是在样本数目趋于无穷大时的渐进理论，在训练样本数目有限的情况下效果不是很好。1995 年，Vapnik 对小样本统计学理论进行了系统化，并在此基础上发展了一种通用的学习方法－支持向量机作为单类分类方法，它能承受特征空间的高维性和表示向量的稀疏性。SVM 在模式识别中的思想是构造一个超平面作为决策平面，使正负之间的空白最大。模式识别的主要任务是构造一个目标函数，使两类模式正确地分开。在线性可分和线性不可分的情况下，可转换为一个典型的二次规划问题。

（7）统计分析方法。

在数据库字段项之间存在两种关系：函数关系（能用函数公式表示的确定性

关系）和相关关系（不能用函数公式表示，但仍是相关确定性关系），对它们的分析可采用统计学方法，即利用统计学原理对数据库中的信息进行分析。可进行常用统计、回归分析、相关分析和差异分析等。

1.1.4 数据挖掘的应用和研究方向

随着 DM 研究的不断深入，DM 技术已逐渐成熟，它的应用也越来越广泛。目前，DM 的应用主要集中在以下几个方面：①金融数据分析；②商业零售数据分析；③电信和网络数据分析；④生物医学和 DNA 数据分析；⑤天文、陆地和海洋地理等科学探测数据分析等；⑥信息检索和分类；⑦Internet 站点访问模式发现；⑧产品产量和质量分析。

数据、DM 任务和 DM 方法的多样性给 DM 提出了许多挑战性课题。DM 语言的设计、高效而有用的 DM 算法和系统的开发、交互的 DM 环境的建立以及应用 DM 技术解决大型应用问题都是 DM 研究人员、系统和应用开发人员所面临的主要问题。DM 是一个新兴的研究领域，许多问题还有待研究。

目前，DM 的研究方向主要有：①DM 的应用研究。早期的 DM 主要集中在帮助企业提高竞争能力。随着 DM 的日益普及，它的应用领域也在不断扩大，由于通用 DM 系统在处理特定应用问题时有其局限性，因此，目前的一种趋势是开发针对特定应用的 DM 系统；②可伸缩的 DM 算法研究。DM 通常是直接面向海量数据库，因此，DM 系统必须能有效地处理海量数据，其算法必须是高效率的、可伸缩的；③DM 系统的交互性。DM 中操作者的适当参与能加速 DM 过程。一方面，交互界面为用户表达要求和策略提供了方便；另一方面，交互界面又把生成的结果传递给用户，由于生成的结果可以多种多样，因此，准确而直观地描述挖掘结果和友好而高效的用户界面一直是研究的重要课题；④DM 语言的标准化研究。标准的 DM 语言或有关方面的标准化工作将有助于 DM 系统的研究和开发，有利于用户学习和使用 DM 系统；⑤DM 的可视化研究。可视化 DM 是从大量数据中发现知识的有效途经。系统研究和开发可视化 DM 技术将有助于推进 DM 作为数据分析的基本工具；⑥对复杂数据类型进行挖掘的新方法研究。目前 DM 系统处理的数据库大多是关系数据库。随着数据库应用范围的日益扩大，规模和功

能的日益完善，数据库中将包含大量复杂的数据类型，甚至出现新的数据库模型，因此，保证 DM 系统能有效地处理此类数据库中的数据是至关重要的；⑦DM 中的隐私保护与信息安全。DM 能从不同的角度、不同的抽象层上看待数据，这将潜在地影响数据的私有性和安全性。随着计算机网络的日益普及，DM 可能导致的非法数据入侵是实际应用中亟待解决的问题之一；⑧Web 挖掘。由于 Web 上存在大量信息，并且 Web 在当今社会扮演越来越重要的角色，因此，Web 挖掘将成为 DM 中一个重要和繁荣的子领域。

1.2 数据挖掘技术在生物信息学中的应用

2003 年 4 月 14 日中午，美国联邦国家人类基因组研究项目负责人弗朗西斯·柯林斯博士隆重宣布："人类基因组序列图绘制成功，人类基因组计划的所有目标全部实现"。序列图的完成使得生物学数据呈现爆炸式增长，PDB（蛋白质数据库）和 GenBank 数据库等均以几何级数的速度增长。Benson 等人经过研究后认为，目前生物学数据库中存储的数据每 13～15 个月增加一倍，其增长速度比著名的摩尔定律还要快 3 个月，大量的其他生物信息数据还在继续快速增长，处理和分析这些数据库中的数据既艰巨又迫切。科学家还将对人类基因组进行更加深入的研究，一方面寻找不同人群之间的基因差异，另一方面破译不同基因的功能以取得更多的数据，为人类战胜疾病、提高生命质量提供更多的参考。国内外各种生物信息数据处理新方法的研究工作正在激烈地展开，其中，DM 技术在生物信息数据处理中的应用研究具有广阔的空间。

1.2.1 生物信息学的定义和研究范围

生物信息学（Bioinformatics）是一个融合了多学科的领域，包括分子生物学（如生物化学、遗传学和结构生物学等）、计算机科学（计算理论、人工智能、机器学习、动态程序设计和 DM 等）、物理化学（热力学和分子建模等）和数学（算法、建模、概率论和统计学等）。它主要是研究生物基因组中信息的获取、加工、储存、分配、分析和解释的一门新兴交叉学科。由于生物信息学涉及的领域宽，

发展异常迅速，很难对生物信息学的研究范围作出明确的界定。目前涉及的主要领域有：

（1）基因组分析。

“基因”这个术语是由德国植物学家 Hans Winkler 提出的。基因组研究的首要目标是获得生物的整套遗传密码，人的遗传密码有 32 亿个碱基，而现在的 DNA 测序仪每个反应只能读取几百到上千个碱基。要得到人的全部遗传密码，首先要把人的基因组打碎，测完一个个小段的序列后再把它们重新拼接起来。基因组大规模测序的每一个环节都与信息分析紧密相关。寻找新的基因是基因组研究的热点之一，使用生物信息学的方法是发现新基因的重要手段。单核苷酸多态性研究（SNP）也是一个重要的研究课题。有的人吸烟喝酒却长寿，也有人自幼就病痛缠身。这是因为他们基因组中存在着差异，这种差异很多表现为基因组上单个碱基的变异，也就是单核苷酸的多态性（SNP）。

（2）蛋白质组学研究。

“蛋白质组”的概念是由 Mark Wilkins 和 Keith Williams 于 1994 年提出。蛋白质组的定义是一个细胞或者一个组织的基因组所表达的全部蛋白质。蛋白质组的研究包括多种手段，如双向凝胶电泳、双向高效层析技术、质谱技术、酵母双杂交技术等。而生物信息学在蛋白质组学的研究中也扮演着极其重要的角色。蛋白质的空间结构预测是生物信息学的一个重要研究领域。蛋白质的结构预测方法大致可分为三类，即比较建模法（Comparative Modeling Method），反向折叠法（Inversethreading）和从头预测法（Ab Initio Prediction）。尤其是从头预测法，需要大规模的计算资源和高效的并行算法。

（3）在基因组水平上研究分子进化。

生物最本质的特征是进化，进化理论是整个生物学的核心。分子方法的引进从分子水平上描述了进化现象和机制，使得进化论变成一种“硬”科学。近几年，随着基因组序列数据的大量增加，对序列差异和进化关系的理解越来越深入。因此，在后基因组时代的分子进化研究是一个非常有前途的研究领域。基因的起源、内含子的起源、外显子的进化等研究将成为后基因组时代的亮点。

（4）比较基因组学研究。

它是研究不同生物之间在基因组结构和组织上的相似和差异的学科。在后基因组时代，完整基因组数据越来越多，有了这些数据，人们就能对若干重大生物学问题进行分析研究，如生命是从哪里起源的?生命是如何进化的?物种是如何起源的?这些重大的问题只有在基因组水平上才能回答。举例来说，鼠和人的基因组大小相似，都含有约 30 亿碱基对，基因的数目也类似且大部分同源。可是鼠和人的差异却如此之大，这是为什么？同样，有的科学家估计不同人种间基因组的差别仅为 0.1%；人猿间差别约为 1%，但他们表型间的差异十分显著。因此，这种差异不仅应从基因、DNA 序列找原因，也应考虑到整个基因组和染色体组织上的差异。这一工作开创了比较基因组学。

（5）药物设计中的生物信息学手段。

计算机药物辅助设计是利用计算化学基本原理，通过模拟药物与受体生物大分子的相互作用，或者通过分析已知药物结构与活性的内在关系，合理设计新型结构先导化合物的药物设计方法。

（6）生物信息学和基因芯片。

生物信息学和基因芯片是生命科学研究领域中的两种方法和技术，两者密切相关，前者促进了后者的研究和应用，而后者则丰富了前者的研究内容。

（7）系统生物学研究。

系统生物学是生命科学皇冠上的明珠，这是因为在基因和蛋白质之间、蛋白质和蛋白质之间都存在着相互作用，这些复杂的相互作用使得它们构成网状的结构。所以，将基因、蛋白质等在系统水平上进行研究才能更好地揭示生物体的功能。过去的生物信息学研究遵循“从序列到结构再到功能”的研究路子，现在正面临一个转变，即应该遵循“从相互作用到复杂系统（网络）再到功能”的研究策略。

1.2.2 生物信息学中的数据挖掘过程

在生物信息数据的整理加工和分析工作中需要用到多种 DM 方法。针对每个具体的任务，根据需要选择特定的数据库，采用不同的挖掘方法，设计不同的挖

掘算法和实现方式。下面简要说明对于特定的项目，DM 的最终实现需要经历的过程。

（1）确定挖掘任务。

首先，必须明确项目的最终目的，分析项目的可行性。生物信息学计算的核心是序列的比较，这包括同一个序列内不同片段的比较以及两个或多个序列的比对。比较的内容从序列的组分变化、寻找特殊的字段到序列间字母的对应。比较的主要目的在于阐明序列之间的同源关系，从已知序列预测新序列的结构和功能，以及蛋白质结构和功能等。

（2）DM 方法设计。

生物信息数据处理的方法从半经验的直观手段到具备较深刻数学背景的复杂算法，跨度很大。对算法的设计或选择主要考虑它的功能和复杂度。生物信息数据量异常庞大，通常我们从数据库中得到这些数据之后，就算经过选择、转化和削减，数据量还是非常惊人。我们在要求算法达到特定功能的同时，应尽量选择一个更加高效的算法。

（3）数据仓库建立及 DM 体系结构的实现。

各种现存的且不断壮大的生物信息数据库是进行 DM 的基础，也是挖掘对象。我们通常选择一种或几种数据库作为数据基础，但仅仅这样是不够的。决策支持需要将来自异种源的数据统一起来，产生高质量的、纯净的、集成的数据，这就需要建立数据仓库（Data Warehouse）。数据仓库技术包括数据清理、数据集成和 OLAP。数据仓库中数据的选择和提取直接影响到 DM 的整体性能。全面的数据处理和数据分析的基础设施将要围绕数据仓库而系统地建立，这包括存取、集成、合并、多个异种数据库的转换、ODBC/OLEDB 连接、Web 访问和服务工具、报表和 OLAP 分析工具。

1.2.3 数据挖掘在生物信息学中的应用和展望

ANN 是生物信息学中广泛使用的机器学习方法之一，也是最早应用于生物学分析领域的技术。目前，ANN 主要用于蛋白质结构和功能预测等领域；DT 是一种简单的诱导型学习系统，使用近似的离散函数对样本进行估计与分类。Salzberg

用 DT 对蛋白质编码基因进行了定位，Selbig 等将其用于蛋白质的结构预测；贝叶斯网络（Bayesian Network，BN）是一种在一组有关的变量中表达概率关联的图示模型，由两部分构成：一部分是贝叶斯网络结构；另一部分是一组条件概率分布。Cai 等用 BN 对 DNA 序列的接合位点进行了建模，Schmidler 等则用它来预测蛋白质的二级结构；GA 主要用于 DNA 片段的拼装，构造分子序列阵列等；隐马尔可夫模型（Hidden Markov Model，HMM）在序列建模、多重阵列和蛋白质结构预测等方面得到了广泛的应用；SVM 在蛋白质折叠识别、基因芯片数据分析和转录起始位点识别等方面有着广泛的应用。

包括 DM 技术在内的机器学习是一种自动的具有人工智能的学习方法，广泛用于解决现实世界中的许多复杂问题。自从将它引入生物信息学领域，机器学习方法帮助加快了分子生物学结构预测、基因定位、基因组学、蛋白质组学等几个重要领域的研究和不断完善。DM 技术在生物信息学的应用将得益于 DM 方法的不断改进与完善。反过来，生物信息学对工具的高要求也将促进 DM 技术的研究进展。随着 DM 技术的进步和生物信息学研究的不断深入，两者将会不断地相互渗透，越来越紧密地结合在一起。

1.3 本书工作

本书的主要工作包括：

（1）针对目前基因功能预测的电子注释方法的可信度较差，准确率、召回率和调和均值均较低，以及基于实验的注释方法费用昂贵的现实问题，提出了一种基于可变精度粗糙集的基因功能预测方法，并进行了模拟数值实验。

（2）常见的癌症、中风、心脏病、糖尿病、自体免疫疾病、忧郁症和哮喘等疾病都是多个基因上的变异位点与环境因子共同作用的结果，单个基因上的变异对疾病的作用是微效的，导致疾病易感性的原因是多基因 SNP 的联合作用。对这些 SNPs 及其单体型的认识将是人类最终揭示复杂性疾病的遗传基础。提出了一种单体型推断和比较方法。通过对四个种群第 21 号染色体单体型的研究，初步判断四个种群之间在 21 号染色体上对疾病的易感程度的差距。

（3）针对线粒体 DNA 突变率较高、限制性片段长度多态性 RFLP（Restriction Fragment Length Polymorphism，RFLP）在基因组中并不普遍存在、短串联重复序列 STR（Short Tandem Repeat，STR) 的突变率较高，提出了采用 Y 染色体的单核苷酸多态性 SNP（Single Nucleotide Polymorphisms，SNP）数据作为研究现代人类进化的有效数据，并进行了模拟数值实验。

（4）蛋白质氨基酸序列可以确定蛋白质的结构，从而进一步确立其亚细胞位置及功能。SOM 程序可以有效预测蛋白质分子的亚细胞位置，但是其数据归一化，距离函数，以及范围函数的选取会明显的影响预测准确率。使用 Batch-Type SOM 算法，在没有采用增益函数的情况下，准确率仍然得到了保证，时间效率也得到了一定的提高。

本书内容按章节安排如下：

第 1 章，综述了数据挖掘技术简介及其在生物信息学中的应用。

第 2 章，提出了一种应用可变精度粗糙集理论为新的生物序列进行功能注释的方法。采用该方法对 GO 数据库中的数据进行 DM 和预测，验证了提出方法的有效性。

第 3 章，提出了一种单体型推断和比较方法，初步判断四个种群之间在 21 号染色体上对疾病的易感程度的差距。

第 4 章，提出了应用 Y 染色体的 SNP 基因型频率数据作为数据来源来研究现代人的种群进化关系，并且应用它构造了种群之间的进化距离矩阵，结论支持被大多数古生物学证据证明的“走出非洲”假说。

第 5 章，使用 Batch-Type SOM 算法，在没有采用增益函数的情况下，预测蛋白质分子的亚细胞位置的准确率仍得到了保证，时间效率也得到了一定的提高。

第 6 章，全面总结了本书的内容，结合本书的工作基础，给出了需要进一步完善的工作和设想。