数据挖掘技术与其在生物信息学中的应用综述

姜浩娜　　杨芬　　刘晓

摘要：本文提纲挈领的阐述了数据挖掘技术及研究热点，并重点阐述了数据挖掘技术在生物信息学中的应用。
关键字：知识发现　数据挖掘　OLAP

引言

无论是商业企业、科研机构或者政府部门，在过去若干年的时间里都积累了海量的、以不同形式存储的数据资料。但当面对越来越多迅速膨胀的超级数据库时，人们却无从着手去理解数据中包含的信息，更难以获得有价值的知识！原有的决策支持系统（DSS）和领导执行系统（EIS）已不能满足需要。数据挖掘概念的提出，使人们有能力克服这些困难，去发掘出蕴藏在数据中的信息和知识。本文总结了目前学术界和工业界的认识和观点，并进行了比较和总结。本文还就与数据挖掘有关的挖掘过程、数据挖掘在生物信息学中的应用等方面阐述了自己的观点。

1. 数据挖掘的概念

本文从探寻知识发现（Knowledge Discovery in Database）和数据挖掘的关系入手理解数据挖掘。数据挖掘与知识发现是存在交叉的两个概念。对这两个概念之间的关系，流行有两种观点：一种观点认为：数据挖掘与知识发现是等同的概念，只不过在不同的领域叫法不同而已。在科研领域，知识发现使用较多，在工程应用领域多称之为数据挖掘。另一种观点认为数据挖掘是知识发现的一个阶段，而且是核心阶段。该观点给出的定义是：**知识发现，就是从大型数据库中的数据中提取人们感兴趣的知识。这些知识是隐含的、事先未知的、潜在有用的信息。**本文更倾向于第二种观点。本文从知识产生的过程这一角度看待知识发现和数据挖掘，得出以下结论：

（1）知识发现是把低级别的数据转化为高级别数据的整个过程。所谓高级别数据，是具有特殊含义的数据。在工程应用中，根据不同的使用阶段和价值，又细分为信息和知识。信息可被理解为有特殊意义的数据；知识则表达为在特定应用领域，通过使用有价值的信息而在人脑中形成的、具有概括和总结特性的认识。知识可表示为概念（concepts），规则（rules），规律（regulations），模式（patterns）等形式。从知识发现的整个过程来看（图1），数据挖掘是知识发现实现从数据到信息和知识转变的关键一步。**数据挖掘是从大量数据中提取可信的、新颖的、有效的模式的高**

级处理过程。

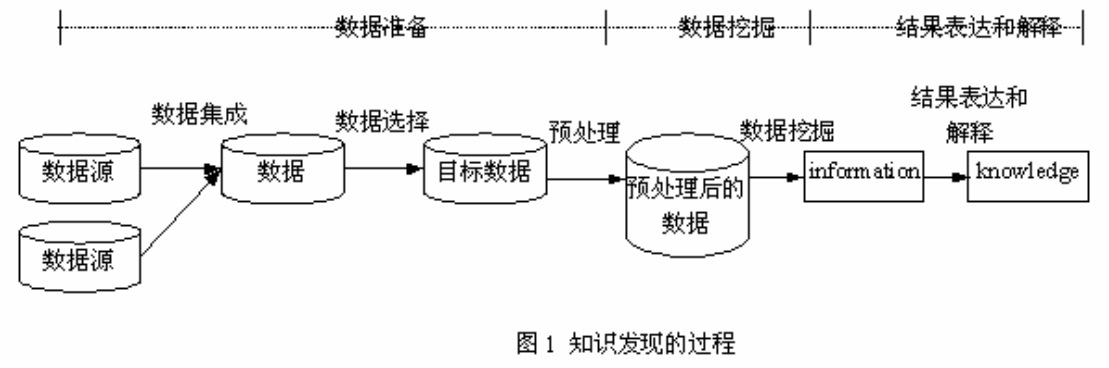

图 1 知识发现的过程

2）如果把知识发现理解为一个过程或系统，数据挖掘是这一过程或系统的一个可自动执行的工具。挖掘算法是数据挖掘重要的组成部分。为解决特定的商业问题，一种或多种算法需要被选择、编译，在适于挖掘的数据环境下实施挖掘任务。从图 1 看出，知识发现是需要人工参与的多环节的过程。除以上谈到的知识发现与数据挖掘的区别外，澄清存在于 OLAP 和数据挖掘之间认识上的混淆会有助于对数据挖掘的理解：Gartner Group 等组织把 OLAP 视为数据挖掘的一部分。数据挖掘包含数据描述和数据建模。OLAP 系统可以提供数据仓库中数据的一般描述。但更多的认识把 OLAP 和数据挖掘当作互不相交的两部分。OLAP 是数据汇总/聚集工具，它帮助简化数据分析。OLAP 的功能基本上是用户参与的汇总和比较（上钻、下钻、旋转、切片、和其他操作）；数据挖掘自动发现隐藏在大量数据中的模式等有价值的知识。图 2 从数据、信息和知识的角度形象地描述出 OLAP 和数据挖掘的逻辑关系。

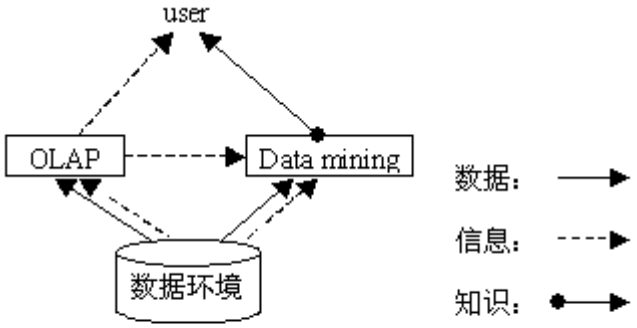

图 2 OLAP 与数据挖掘的关系

另一点，OLAP 大多是限于数据仓库中的数据。数据挖掘既可以分析现存的、比数据仓库提供的汇总数据粒度更细的数据，也可以分析事务的、文本的、空间的和多媒体数据。

2. 数据挖掘分类和挖掘步骤

2.1. 数据挖掘分类

数据挖掘涉及的学科领域和方法很多，有人工智能、数据统计、可视化、并行计算等。数据挖掘有多种分类方法。

2.1.1. 根据挖掘任务

可分为分类模型发现、聚类、关联规则发现、序列分析、偏差分析、数据可视化等。

（1） 分类（Classification）

其旨在生成一个分类函数或分类模型，该模型能把数据库中的数据项映射到给定类别中的某一个。既可以用此模型分析已有的数据，也可以用它来预测未来的数据。

（2） 聚集（Clustering）

聚集是对记录分组，把相似的记录在一个聚集里。聚集和分类的区别是聚集不依赖于预先定义好的类，不需要训练集。

（3） 数据可视化（Description and Visualization）

数据可视化严格地讲不是一个单独的数据挖掘任务，它被用来支持其他挖掘任务。可视化是采用图形、图表等易于理解的方式表达数据挖掘结果。

（4） 关联规则（Affinity grouping or association rules）

关联规则是寻找数据库中值的相关性，主要是寻找在同一个事件中出现的不同项的相关性，比如在一次购买活动中所买不同商品的相关性。

（5） 序列分析（Sequence Analysis）

序列模式分析同样也是试图找出数据之间的联系。但它的侧重点在于分析数据之间前后（因果）关系，因此对数据往往要求引入时间属性。序列模式分析非常适于寻找事物的发生趋势或重复性模式。

（6） 偏差分析（Deviation Analysis）

是用来发现与正常情况不同的异常和变化，并进一步分析这种变化是否是有意的诈骗行为，还是正常的变化。如果是异常行为，则提示预防措施；如果是正常的变化，那么就需要更新数据库记录。

2.1.2. 根据挖掘对象

可分为关系数据库、面向对象数据库、空间数据库、时态数据库、文本数据源、多媒体数据库、异质数据库以及环球网 Web。

2.1.3. 根据挖掘方法

可粗略地分为：机器学习方法、统计方法、神经网络方法、决策树、可视化、最近邻技术等。在

机器学习中，可细分为归纳学习方法(决策树、规则归纳等)、基于范例学习、遗传算法等；在统计方法中，可细分为:回归分析(多元回归、自回归等)、判别分析(贝叶斯判别、费歇尔判别、非参数判别等)、聚类分析(系统聚类、动态聚类等)、探索性分析（主元分析法、相关分析法等）等。

2.2. 数据挖掘步骤

数据挖掘包括商业需求、大量的数据和挖掘算法三部分。商业需求是真正的数据挖掘前期要明确的工作。挖掘算法是目前研究的热点之一，主要围绕采用新的挖掘算法解决特定商业问题和对挖掘算法的改进上。由挖掘算法可形成挖掘工具。它的一般步骤是：

（1）分析问题：源数据数据库必须经过评估确认其是否符合数据挖掘标准。以决定预期结果，也就选择了这项工作的最优算法。

（2）提取、清洗和校验数据：提取的数据放在一个结构上与数据模型兼容的数据库中。以统一的格式清洗那些不一致、不兼容的数据。一旦提取和清理数据后，浏览所创建的模型，以确保所有的数据都已经存在并且完整。

（3）创建和调试模型：将算法应用于模型后产生一个结构。浏览所产生的结构中数据，确认它对于源数据中"事实"的准确代表性，这是很重要的一点。虽然可能无法对每一个细节做到这一点，但是通过查看生成的模型，就可能发现重要的特征。

（4）查询数据挖掘模型的数据：一旦建立模型，该数据就可用于决策支持了。在微软的数据挖掘解决方案中，该过程通常使用 VB 或 ASP 通过 OLE DB for Data Mining Provider 编写前端查询程序。

（5）维护数据挖掘模型：数据模型建立好后，初始数据的特征，如有效性，可能发生改变。一些信息的改变会对精度产生很大的影响，因为它的变化影响作为基础的原始模型的性质。因而，维护数据挖掘模型是非常重要的环节。

3. 数据挖掘的研究热点

目前，数据挖掘的研究热点围绕挖掘算法展开。数据挖掘是综合了机器学习、人工智能、数据统计等学科的研究领域。随着数据挖掘工具在实际应用中的迅速增长，相关学科成熟的算法实现不断地加入到数据挖掘中来；挖掘算法的研究还包括对现有挖掘算法的优化和改进，比如使用全局搜索算法优化神经网络学习过程。除此以外，数据挖掘还有以下研究热点：

3.1.数据挖掘原语

我们把数据挖掘原语可以理解为有效辅助数据挖掘实施知识发现功能的思想和做法，是知识发现的辅助工具。数据挖掘原语的研究是为了支持有效的知识发现，为了让用户更加易于理解挖掘出来的知识。用户能够通过数据挖掘原语与数据挖掘系统通信，从不同的角度和深度审查发现结果，并指导挖掘过程。这组原语包括数据库说明的部分或用户感兴趣的数据集、要挖掘的知识类型、用于指导挖掘过程的背景知识、模式评估以及度量和如何显示所发现的知识等等。

3.2.数据挖掘语言及标准

数据挖掘行业是高度分散的，公司和研究机构独立开发各自的数据挖掘系统和平台，没有形成开放性的标准；同时数据挖掘本身是一门多学科综合跨度非常大的技术，这些造成了数据挖掘在通用性方面存在一系列问题 ：①各种数据挖掘问题及挖掘方法基于不同的模型和技术，彼此互相孤立，联系很少；②缺少简明精确的问题描述方法，挖掘的语义通常是由实现方法决定的；③数据挖掘系统仅提供孤立的知识发现功能，难于嵌入大型应用；④数据挖掘引擎与数据库系统是松散耦合的。数据挖掘语言和标准的开发有望解决上述问题。

目前，已实现的数据挖掘语言有 DMQL、MSQL 和 MINE RULE 等；数据挖掘语言标准有 PMML（DMG：数据挖掘组织协会）和 OLE DB for Data Mining（微软）。

3.3.数据挖掘系统

知识发现是一个有机的整体，各个部分之间有着密切的关系。我们称围绕某一数据挖掘任务的知识发现过程为数据挖掘系统。应该说所有的算法是为某一个挖掘系统服务的。数据挖掘系统的研究是为了建立科学的系统结构，利于挖掘算法的重用、嵌入，利于算法与系统其他模块有机结合。图 3 是一个挖掘系统的原型结构。

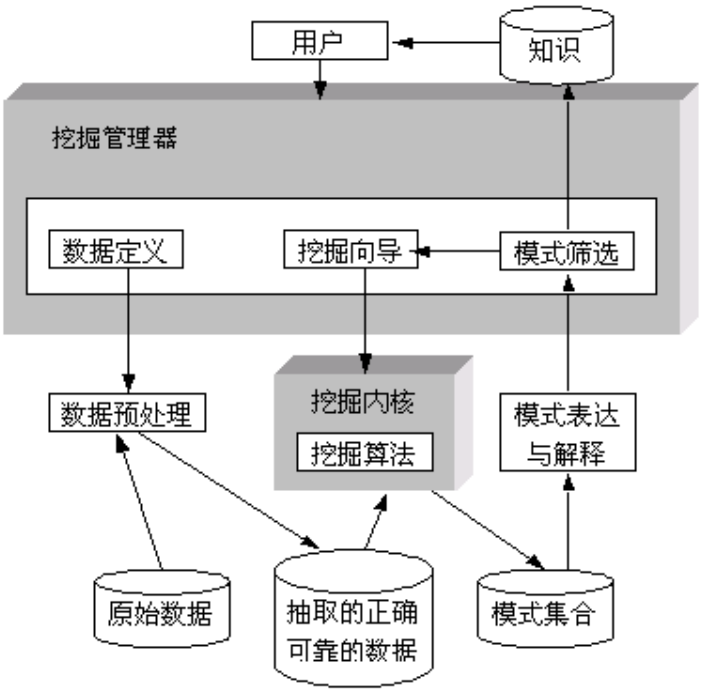

图 3 一个数据挖掘系统原型

4. 数据挖掘在生物信息学中的应用

生物信息学是一门新兴的交叉学科。人类基因组计划的启动和实施使得核酸、蛋白质数据迅速增长,如何从海量数据中获取有效信息成为生物信息学迫切要解决的问题。数据挖掘与生物信息学有很好的结合点,在生物信息学领域的应用潜力日益受到人们的重视。研究证明数据挖掘技术是生物信息处理的强有力工具。目前数据挖掘在生物信息领域的研究重点主要表现在以下几个方面:

(1) 数据清理,数据集成,异种、分布式数据库的语义集成。

许多国家和研究组织都建立了生物序列数据库、蛋白质结构和功能数据库,为人们提供了丰富的信息。但是这些数据分散,且存储介质多样,在同一数据库中存在着大量具有重复信息的序列及一些高度相似的数据,造成数据冗余。因此对这种异构的和广泛分布的数据库的语义集成就成为一项重要任务。数据挖掘中的数据清理、数据集成方法有助于该问题的解决。

(2) DNA 序列相似搜索和比对。

为识别一个新发现的基因和一个已知基因家族之间的进化关系,确定他们的同源性或相似性,通常需要序列比对,找出它们之间的最大匹配,从而定量给出其相似程度。由于序列数据是非数字的,其内部不同种类核苷酸之间的精确交叉扮演着重要的角色。因此探索高效的搜索和比对算法在序列分析中非常重要。

(3) 基因组特征及同时出现的基因序列的分析。

对于基因家族的成组序列来说,必须阐明多个序列之间的关系,才能揭示整个基因家族的特征。多序列比对在识别一组相关序列中有重要生物意义。多比对算法的计算量可观,为降低算法复杂性,必须研究有实用价值的比对算法。利用关联规则、聚类分析有助于发现一组序列之间的差异以及相似性关系,以便对一个基因家族的特征有基本了解。另外在生物医学研究上,人们发现疾病的产生大多数是由多基因决定的,利用关联规则分析帮助确定在目标(疾病) 样本中同时出现的基因种类。

(4) 路径分析:发现在不同阶段的致病因。

引起一种疾病的基因不止一个,不同基因在疾病的不同阶段发挥作用。利用路径分析、演变分析等找到在不同阶段的致病遗传基因序列,可开发不同阶段的治疗药物,从而取得更有效的治疗效果。

(5) 生物数据可视化和可视的数据挖掘。

由于生物数据的复杂性和高维性,既不能以数字公式表示,也不能以逻辑公式表示,可借助各种可视化工具以图、树、方体、链的形式展现其复杂结构和序列模式。常用的生物数据可视化工具有语义镜技术、信息壁技术、基因调控网格等 。同时,将经过数据挖掘工具得到的数据结果也以图形、图像的形式展现给用户,便于用户寻找数据间规律和关系。

(6) 生物文献的挖掘。

Internet 上生物文献日益增多,人们通过搜索引擎获取相关信息,但检索结果数目巨大,准确率不高,而且一般搜索结果只能给出标题和摘要,不能给出文章总的关键词句。特别是生物文献数据中,大多数蛋白质名称都是复合词,比较复杂,有的却采用普通的词汇命名与其常用词义相混导致引擎搜索错误,甚至在一篇文献中同一种蛋白质有好几种命名出现,增加了搜索困难。利用路径遍历模式、链接分析、自然语言处理等技术寻找文献中关键词如蛋白质名称,或捕捉上下文关系,可提高检索速度和准确率 。

(7) 基于隐私保护的数据挖掘。数据挖掘技术为生物工作者提供了有效工具的同时也引发了隐私保护问题。比如研究单位的保密实验数据,个人的医疗诊断记录、病史记录都有可能被误用。通过在数据挖掘过程中使用限制数据访问,模糊数据,减少不必要分组,有目的增加噪声数据等方法来达到保护隐私的目的 。目前在该领域的研究尚处于起步阶段。

随着多年的研究与发展,已有很多数据挖掘、机器学习系统和工具用于生物信息处理。一般的数据挖掘分析系统有: SAS Enterprise Miner , IBM Intelligent Miner ,SGIMinSet 等。一些专用的综合软件包在生物信息处理中发挥了巨大作用。GCG(Genetics Computer Group) 主要用于

核酸序列分析和蛋白质序列分析。Staden 是 DNA 和蛋白质序列分析的软件包。此外还有用于大规模测序的 Sequencher ,用于快速克隆的 VectorNTI 等。GeneMine 是由 Molecular Application Group 开发的生物信息学数据挖掘系统,该系统可以用于生物信息数据的过滤、计算和聚类操作,并支持进一步的综合分析和可视化。目前世界数据库巨头 ORACL E , IBM 纷纷将生物信息挖掘工具分别嵌入至 ORACL E 9i ,DB2 中,大大提高了生物数据的安全性和分析的准确性。

作者介绍:

姜浩娜　深圳福田人民医院信息科　　工程师　　中山大学　硕士

TEL 13714546272　　email：alala_1997@263.net

杨　芬　深圳福田人民医院信息科　　主任　　同济大学　硕士

TEL 13688809668　　email：ftyangfen@sina.com

刘　晓　深圳福田人民医院信息科　　工程师