计算机理论基础与应用丛书

模糊关联规则的研究与应用

陆建江　张亚非　宋自林　/著

科学出版社
www.sciencep.com

计算机理论基础与应用丛书

模糊关联规则的研究与应用

陆建江　张亚非　宋自林　著

科学出版社

北京

内 容 简 介

关联规则发现是数据挖掘中最重要的任务之一，它的目标是发现数据中属性之间有趣的关联。数量型关联规则是一种重要的关联规则类型，用来发现数量型属性之间的关联。本书应用模糊集来软化属性论域的划分边界，并系统地介绍数量型属性的模糊关联规则及其应用。主要内容包括：数量型属性的模糊关联规则及其挖掘算法；集合值和区间值关系数据库上模糊关联规则及其挖掘算法；加权模糊关联规则及其挖掘算法；模糊关联规则的并行挖掘算法；模糊关联规则的增量更新；关注模糊关联规则的挖掘算法；模糊关联规则在分类和预测中的应用等方面。

本书可作为高等院校计算机专业研究生的教材，也可作为相关领域学生的参考书。

图书在版编目(CIP)数据

模糊关联规则的研究与应用/陆建江，张亚非，宋自林著. —北京：科学出版社，2008

（计算机理论基础与应用丛书）

ISBN 978-7-03-020556-8

Ⅰ. 模… Ⅱ. ①陆…②张…③宋… Ⅲ. 模糊集-研究 Ⅳ. O159

中国版本图书馆 CIP 数据核字（2007）第 189022 号

责任编辑：张海娜/责任校对：陈玉凤

责任印制：刘士平/封面设计：王 浩

科 学 出 版 社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

新 蕾 印 刷 厂 印刷

科学出版社发行 各地新华书店经销

*

2008 年 1 月第 一 版 开本：B5（720×1000）

2008 年 1 月第一次印刷 印张：9 1/2

印数：1—3 000 字数：180 000

定价：28.00 元

（如有印装质量问题，我社负责调换（明辉））

前　言

在信息爆炸的时代，信息过量几乎成为人人都需要面对的问题。人们迫切需要一种新的强有力的数据分析工具来自动和智能地将待处理的数据转化为有用的信息和知识。数据挖掘是从大量的、不完全的、有噪声的、模糊的和随机的数据中，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。关联规则发现是数据挖掘中最重要的任务之一，它的目标是发现数据中属性之间有趣的关联。数量型关联规则是一种重要的关联规则类型，用来发现数量型属性之间的关联。

本书是一本全面、系统地介绍数量型属性的模糊关联规则及其应用的学术专著。作者从 1998 年起就开展数据挖掘特别是数量型属性的模糊关联规则的研究工作，全书的内容是作者 9 年来在这方面的研究成果，也是作者在解放军理工大学的博士论文与在东南大学的博士后报告主要内容的汇总。本书深入浅出，并主要通过图表辅助等方式循序渐进地介绍模糊关联规则的知识，含有大量的应用实例，这些应用实例能引导渴望了解和有志于研究模糊关联规则技术的读者逐步、系统地熟悉和了解整个模糊关联规则涉及的基础知识和面临的关键问题。

本书的内容共分九章。第一章介绍数据挖掘的基本概念、基本内容和基本过程，重点介绍布尔型关联规则和数量型关联规则的发展现状，分析了目前数量型关联规则挖掘算法存在的问题。第二章应用模糊集来软化属性论域的划分边界，并系统地提出了数量型属性的模糊关联规则及其挖掘算法。第三章提出集合值和区间值关系数据库上模糊关联规则及其挖掘算法。第四章提出加权模糊关联规则及其挖掘算法。第五章提出模糊关联规则的并行挖掘算法。第六章提出模糊关联规则的增量更新。第七章提出关注模糊关联规则的挖掘算法。第八章提出基于模糊关联规则在分类中的应用。第九章提出基于模糊关联规则在预测中的应用。

陆建江副教授、张亚非教授、宋自林教授经过近一年的艰苦努力完成了全书的编写工作。另外，周波、李言辉、康达周、苗壮等同学参与了全书的校对工作，在此不胜感谢。全书每一章的内容组织和细节都经过多次讨论和修改才定稿，力求深入浅出，让读者轻松掌握相关的知识。尽管每一节、每一句、每篇参考文献、甚至每个标点我们都精心检查，但难免还存在一些缺点和错误，殷切希望广大读者批评指正。我们希望本书的出版能够对数据挖掘和知识发现相关领域

的研究人员有所裨益，本书可作为高等院校计算机专业研究生的教材，也可作为相关领域学生的参考书。

十分感谢科学出版社的同志为本书的出版所做的工作。

目　录

第一章 绪 论

1.1 数据挖掘概述

近十几年来，人们利用信息技术生产和搜集数据的能力大幅度提高，越来越多的数据库被用于商业管理、政府办公、科学研究和工程开发等，并且这一势头仍将持续发展下去。于是，一个新的挑战被提了出来：在这信息爆炸的时代，信息过量几乎成为人人都需要面对的问题，仅仅依靠数据库管理系统的查询检索机制和统计学分析方法已经远远不能满足现实的需要，它迫切要求开发一种新的强有力的数据分析工具自动和智能地将待处理的数据转化为有用的信息和知识。因此，面对“人类正被数据淹没，却饥渴于知识”的挑战，数据挖掘和知识发现（DMKD）技术应运而生，并得以蓬勃发展，越来越显示出其强大的生命力。

1.1.1 数据挖掘概念

Fayyad 等人将数据库中的知识发现（KDD）定义为“从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程”（Fayyad et al.，1996）。实际上，KDD 的目标在于发现未知的、有用的并且简洁的模式，它是一个交叉的研究领域，吸引了许多机器学习、模式识别、数据库、统计学、人工智能、专家系统、图形理论和数据可视化等相关领域的研究者。KDD 系统通常使用来自这些领域的方法、算法和技术。

KDD 过程是一个使用数据挖掘技术，根据一些特定的度量方法和阈值提取有意义知识的交互和迭代式多阶段过程，如图 1-1 所示。KDD 的主要步骤包括：

（1）数据准备。它包括 3 个子步骤：数据集成、数据选择和数据预处理。数据集成将多文件或多数据库运行环境中的数据进行合并处理，解决语义模糊性、处理数据中的遗漏和清洗脏数据等。数据选择的目的是辨别出需要分析的数据集合，缩小处理范围，提高数据挖掘的质量。数据预处理是为了克服目前数据挖掘工具的局限性。

（2）数据挖掘。该阶段先要决定如何产生假设，是让数据挖掘系统自动为用户产生假设，还是用户自己对于数据库中可能包含的知识提出假设。前一种称为发现的数据挖掘，后一种称为验证型的数据挖掘。然后选择合适的工具进行实际的挖掘操作，发现有用的模式或知识。

（3）结果表达和解释。根据最终用户的决策目的对提取的信息进行分析，把

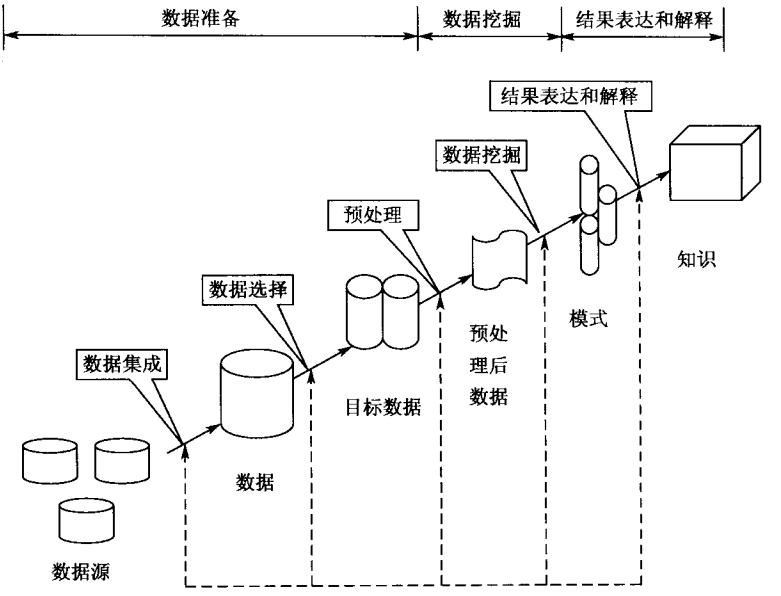

图 1-1 KDD 过程

最有价值的信息区分出来，并且通过决策支持工具提交给决策者，因此这一步骤任务不仅是把结果表达出来，还要对信息进行过滤处理，如果不能令决策者满意，需要重复以上的数据挖掘过程。

显然，数据挖掘是 KDD 过程的一个核心步骤。然而，在产业界、媒体和数据库业界，术语“数据挖掘”比术语“数据库中的知识发现”使用更普遍、更流行。

数据挖掘可以认为是信息技术自然演变的结果。在数据库业界，数据挖掘的进化经历了四个阶段：数据搜集、数据访问、数据仓库和决策支持。数据挖掘的进化历程见表 1-1。

表 1-1 数据挖掘的进化历程

进化阶段	支持技术	产品厂家	产品特点
数据搜集 (20 世纪 60 年代)	计算机、磁带和磁盘	IBM，CDC	提供历史性的、静态的数据信息
数据访问 (20 世纪 80 年代)	关系数据库、结构化查询语言、ODBC	Oracle，Sybase，IBM，Informix，Microsoft	在记录级提供历史性的、动态的数据信息
数据仓库； 决策支持 (20 世纪 90 年代)	联机分析处理、多维数据库、数据仓库	Pilot，Comshare，Arbor，Cognos，Microstrategy	在各种层次上提供回溯的、动态的数据信息
数据挖掘	高级算法、多处理器计算机、海量数据库	Pilot，Lockheed，IBM，SGI，其他初创公司	提供预测性的信息

数据挖掘定义为是从大量的、不完全的、有噪声的、模糊的和随机的数据中，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。还有很多和它相近似的术语，如知识提取、数据/模式分析、数据考古以及数据捕捞等。人们把原始数据看做形成知识的源泉，就像从矿石中采矿一样。原始数据可以是结构化的，如关系数据库中的数据；半结构化的，如 XML 表示的数据；无结构化的，如文本、图形和图像数据；甚至可以是分布在网络上的异构型数据。发现知识的方法可以是数学的，也可以是非数学的；可以是演绎的，也可以是归纳的。发现了的知识可以被用于信息管理、查询优化、决策支持和过程控制等，还可以用于数据自身的维护。

1.1.2 数据挖掘的任务

数据挖掘有描述和预测两个高层目标。描述性的数据挖掘任务是刻画数据库中数据的一般特征。预测性的数据挖掘任务是建立一个或一组模型用于对新数据进行预测。数据挖掘的任务主要包括以下六个方面（Han et al.，2000a）。

1. 类/概念描述：特征化与区分

数据可以与类或概念相关联。用汇总的、简洁的和精确的方式描述每个类或概念可能是有用的。这种类或概念的描述称为“类/概念描述”。这种描述可以通过下述方法得到：

（1）数据特征化，一般地汇总目标类的数据。

（2）数据区分，把目标类与一个或多个比较类进行比较。

（3）数据特征化和数据区分两者都有。

数据特征化是目标类数据的一般特征或特性的汇总。通常，用户指定类的数据通过数据库查询收集。例如，为研究上一年销售增加 10%的软件产品的特征，就可以通过执行一个 SQL 查询收集关于这些产品的数据。

有许多有效的方法将数据特征化。例如，基于数据立方体的联机分析处理（OLAP）上卷操作可用来执行用户控制的、沿着指定维的数据汇总。一种面向属性的归纳技术可以用来进行数据的泛化与特征化，而无须一步步地与用户交互。数据特征的输出可以用多种形式提供。包括饼图、条图、直方图、曲线、多维数据立方体以及交叉表在内的多维表。结果描述也可以用泛化关系或规则形式提供。

数据区分是将目标类对象的一般特性与一个或多个对比类对象的一般特性进行比较。目标类与对比类由用户指定，而对应的数据通过数据库查询检索。例如，你可能希望将上一年销售额增加 10%的软件产品与同一时期销售额下降 30%的软件产品进行比较。尽管数据区分还应包括用来区分目标类与对比类的比

较度量，用于数据区分的方法与用于数据特征化的类似。输出的形式也相似，用规则表示的区分描述，称为区分规则。

2. 关联分析

关联分析是发现关联规则，这些规则展示属性-值频繁地在给定数据集中一起出现的条件。一般地，关联规则有形式“$X \Rightarrow Y$”，即“$A_1 \wedge \cdots \wedge A_m \Rightarrow B_1 \wedge \cdots \wedge B_n$”，这里 A_i，$i \in \{1,\cdots,m\}$ 和 B_j，$j \in \{1,\cdots,n\}$ 是属性-值对。关联规则“$X \Rightarrow Y$”可解释为“满足条件 X 的数据库记录很有可能也满足条件 Y”。例如，购买计算机也趋向于同时购买财务管理软件可以用以下布尔型关联规则表示：

$\text{computer} \Rightarrow \text{financial\_management\_software}(\text{Support}=3\%, \text{Confidence}=70\%)$

其含义是：这条关联规则支持率为 3%，信任度为 70%，即购买计算机同时也购买财务管理软件的可能性为 70%，且这样的顾客占顾客总数的 3%。除了布尔型关联规则，另外一种重要的关联规则是数量型关联规则，例如，一个数据挖掘系统在商场的数据中可能发现如下形式的数量型关联规则：

$\text{age}(X,\text{“20} \sim \text{29”}) \wedge \text{income}(X,\text{“20k} \sim \text{30k”}) \Rightarrow \text{buys}(X,\text{“CDplayer”})$

$(\text{Support} = 2\%, \text{Confidence} = 60\%)$

其含义是：年龄在 20～29 之间，且收入在 20k～30k 之间的顾客购买 CD 播放机的可能性为 60%，且这样的顾客占顾客总数的 2%。

3. 分类与预测

分类是找出描述并区分数据类或概念的模型或函数（也常称作分类器），以便能够使用模型预测类标记未知的对象类。它以一组训练数据的分析为基础，每个目标数据有一个已知的唯一的类标签。理想情况下，给定目标类的模型将能描述该类中的所有目标数据，但不能描述其余类中的目标数据。然而在现实世界中往往不可能导出如此理想的模型。一般地说，机器学习文献把这种不可能性归于数据中出现的噪声或孤立数据，以及为了避免模型去过分适合此类数据。机器学习观点还倾向于目标数据是可唯一分类的，也就是说，每个训练目标数据明确地属于一个类且仅属于一个类。相应地，数据挖掘的分类观点认为，尽管由于噪声和为了避免“过分的适合”使获得理想模型成为不可能，但妨碍生成理想模型的主要因素还在于大型数据库中数据的多样性和大量性。正由于数据的多样性，假定所有目标数据可以被唯一地分类是没有理由的。更可能的是，一个给定的目标数据可能属于多个类，特别是当数据已被汇总到更高的抽象层次时（在处理现实世界的大型数据库时经常遇到这种情况）。因此数据挖掘的分类观点认为：得到理想的数据模型是不可能的，每个模型最终将覆盖它所代表的类中的大多数目标数据，同时最大限度地把此类的特性与其他类的特性区分开来。此外，当用这些

模型去识别一个类标签未知的目标数据时，分类模型将返回一个类的概率分布，而不是一个唯一的类预测。这个概率分布可使用户看到该目标数据属于每一个类的预测概率。

分类器的构造方法有统计方法、机器学习方法和神经网络方法等。统计方法包括贝叶斯法和非参数法（近邻学习或基于事例的学习），对应的知识表示则为判别函数和原型事例。机器学习方法包括决策树法和规则归纳法，前者对应的表示为决策树或判别树，后者则一般为产生式规则。神经网络方法主要是 BP 算法，它的模型表示是前向反馈神经网络模型（由代表神经元的节点和代表连接权值的边组成的一种体系结构），BP 算法本质上是一种非线性判别函数。另外，最近又兴起了两种新的分类方法：粗糙集方法，其知识表示是产生式规则；Boosting 分类方法，其知识表示是一个分类函数。

当分类不是用来预言类标签，而是用来预测空缺的或不知道的数据值，这个过程称为预测。在分类与预测之前首先要进行相关分析，它试图确定那些对分类与预测没有作用的属性。这些属性应当被排除在外。

4. 聚类

与分类与预测不同，聚类分析的对象是不带有类标签的目标数据。一般地说，类标签不出现在训练数据中仅仅是因为一开始不知道它们。聚类能自动生成类标签。目标数据根据类内的相似性最大、类外的相似性最小的原则进行聚集。每个形成的聚类可被看成一个目标类，从目标类中可得出规则。

5. 孤立点分析

数据库中可能包含一些数据对象，它们与数据的一般行为或模型不一致。这些数据对象是孤立点。大部分数据挖掘方法将孤立点视为噪声或异常而丢弃。然而，在一些应用中（如欺骗检测），罕见的事件可能比正常出现的那些更有趣。孤立点数据分析称为孤立点挖掘。孤立点可以使用统计试验检测。它假定一个数据分布或概率模型，并使用距离度量，到其他聚类的距离很大的对象被视为孤立点。基于偏差的方法通过考察一群对象主要特征上的差别识别孤立点，而不是使用统计或距离度量。

6. 演变分析

数据演变分析描述随时间变化的对象的行为规律或趋势，并对其建模。尽管这可能包括时间相关数据的特征化、区分、关联、分类或聚类，这种分析的不同特点包括时间序列数据分析、序列或周期性模式匹配和基于相似性的数据分析。

数据挖掘是一个交叉学科领域，受多个学科的影响，根据不同的标准有多种

分类法。根据发现知识的种类分类，可分为概念/类的描述、关联规则发现、分类与预测模型发现、聚类分析发现、孤立点分析发现、演变分析发现等；根据挖掘知识的抽象层次分类，可分为原始层次的数据挖掘、高层次的数据挖掘和多层次的数据挖掘等；根据挖掘的数据库分类，可分为关系型、事务型、面向对象型、主动型、空间型、时间型、文本型、多媒体、异质数据库和遗产数据库上的数据挖掘等；也可以根据所用的技术进行分类，如常用的挖掘技术是数据库技术、人工神经网络、决策树、遗传算法、最近邻技术、规则归纳和可视化技术等。

1.1.3 数据挖掘的方法和技术

数据挖掘的典型方法和技术主要有：归纳学习方法、仿生物技术、公式发现、统计分析方法、模糊论方法及可视化技术等。

归纳学习方法是目前重点研究的方向，研究成果较多。从采用的技术上看，可分为两大类：信息论方法和集合论方法。信息论方法是利用信息论的原理建立决策树。知识表示形式为决策树，故常称为决策树方法。该类方法的实用效果好，影响较大。典型方法的有 ID3 方法、C4.5 方法和 IBLE 方法。集合论方法是开展较早的方法，近年来，由于粗糙集理论的发展使集合论方法得以迅速发展。这类方法包括：覆盖正例排斥反例的方法如 AQ、概念树方法和粗糙集方法。

仿生物技术典型的方法是神经网络方法和遗传算法。这两类方法已经形成了独立的研究体系，它们在数据挖掘中发挥了巨大作用。神经网络模型主要包括前馈式网络、反馈式网络和自组织网络。

公式发现方法在工程和科学数据库中对若干数据项进行一定的数学运算，求得相应的数学公式。典型的系统有：物理定律发现系统 BACON、经验公式发现系统 FDD。

统计分析方法利用统计学原理对数据库中的数据进行分析。包括相关分析和回归分析、差异分析、聚类分析及判别分析等。

模糊论方法利用模糊集合理论对实际问题进行模糊评判、模糊决策、模糊模式识别和模糊聚类分析等。由于模糊性是客观的存在，而且系统的复杂性愈高，精确化能力愈低，这就意味着模糊性愈强，这是 Zadeh 总结的互克性原理。以上提到的模糊论方法在实际应用中都取得了较好的效果。

可视化数据分析技术拓宽了传统的图表功能，使用户对数据的剖析更清楚。例如，把数据库中多维数据变成多种图形，这对于揭示数据中的状况、内在本质以及规律性起到很强的作用。

1.1.4 数据挖掘工具的评价标准

如何选择满足自己需要的数据挖掘工具呢？评价一个数据挖掘工具，需要从以下几个方面来考虑。

1. 产生的模式种类的多少

数据挖掘系统能生成成千上万的模式与规则，生成的模式仅仅只有一部分对给定的用户有用。这就引出了一些重要的数据挖掘问题：什么样的模式有意义？一个数据挖掘系统能否生成所有有意义的模式？一个数据挖掘系统能否仅仅生成有意义的模式？

模式是否有意义包含以下几个方面：①能否容易被人理解；②在新数据或试验数据上有一定程度的正确性；③潜在的有用性；④新颖性。如果一个模式能证明一个用户想证实的假设，那么模式也是有意义的。一个有意义的模式代表知识。

存在许多判断模式是否有意义的客观度量。这些客观度量依赖于发现模式的结构和关于它们的统计量。对于形如“$X \Rightarrow Y$”的关联规则，一种客观度量是规则的支持率，规则的支持率代表满足给定规则样本数据的百分率。关联规则的另一个客观度量是规则的信任度，它评估被挖掘规则的信任程度，即假定模式 X 是真时，Y 也是真的条件概率。支持率与信任度的定义如下：

$$\text{Support}(X \Rightarrow Y) = \text{Prob}\{X \cup Y\}$$

$$\text{Confidence}(X \Rightarrow Y) = \text{Prob}\{Y \mid X\}$$

一般地说，每一个有意义的度量都与一个阈值联系在一起，阈值由用户控制。例如，不满足信任度阈值 50% 的规则认为是没意义的。低于阈值的规则可能反映噪声、异常和少数情形，也可能是值小。

尽管客观度量能够帮助识别有意义的模式，但仅有客观度量还是不够的，还需要能反映特殊用户需要和兴趣的主观度量。许多对客观度量有意义的模式也许只代表公共知识，因此实际上可能没有意义。主观度量以用户在数据上的信念为基础，如果发现的模式是意外的（与用户的想法相反）或能给用户的行动提供策略性的信息，则发现的模式是有意义的。后者的模式称为可行动的。

“一个数据挖掘系统能否生成所有有意义的模式”指的是一个数据挖掘算法的完备性。如果一个数据挖掘系统生成所有可能的模式，这个系统是不现实的和无效的。实际上，应当根据用户提供的限制和信任度对搜索聚焦。对于某些数据挖掘任务，保证数据挖掘算法的完备性是可达到的。关联规则的挖掘即是一个例子，在那里，有意义的度量能确保挖掘算法的完备性。

最后一个问题“一个数据挖掘系统能否仅仅生成有意义的模式”是数据挖掘

的一个优化问题。能否仅仅生成有意义的模式是数据挖掘系统强有力的心愿。这更是用户的心愿，因为用户可以无需再去搜寻生成的所有模式来确定模式是否真正有意义。这个优化问题目前仍是数据挖掘的挑战性问题。

对用户来说，模式有意义的度量是发现有用模式的必要条件。在数据挖掘步骤后，根据这种度量对发现模式的意义进行排序，或过滤没有意义的模式。更重要的是，这种度量能指导和限制发现过程，并通过剪除不满足预先给定有意义阈值的模式空间子集来提高搜索速度。

2. 解决复杂问题的能力

数据量的增大，对模式精细度、准确度要求的增高都会导致问题复杂性的增大。数据挖掘系统可以提供下列方法解决复杂问题：

多种类别模式的结合使用有助于发现有用的模式，降低问题复杂性。例如，首先用聚类的方法把数据分组，然后再在各个组上挖掘预测性的模式，将会比单纯在整个数据集上进行操作更有效、准确度更高。

多种算法，特别是与分类有关的模式，可以用不同的算法来实现，各有各的优缺点，适用于不同的需求和环境。数据挖掘系统提供多种途径产生同种模式，将更有能力解决复杂问题。验证方法在评估模式时，有多种可能的验证方法。比较成熟的方法有 N 层交叉验证。

有用的数据和有意义的模式通常隐藏在大量的数据中。有些数据是冗余的，有些数据是完全无关的。而这些数据项的存在会影响到有价值的模式的发现。数据挖掘系统一个很重要的功能就是能够处理数据复杂性，提供工具，选择正确的数据项和转换数据值。

可视化工具提供直观、简洁的机制表示大量的信息。这有助于定位重要的数据，评价模式的质量，从而减少建模的复杂性。扩展性为了更有效地提高处理大量数据的效率，数据挖掘系统的扩展性十分重要。需要了解的是：数据挖掘系统能否充分利用硬件资源？是否支持并行计算？算法本身设计为了并行的或利用了DBMS 的并行性能？支持哪种并行计算机，SMP 服务器还是 MPP 服务器？当处理器的数量增加时，计算规模是否相应增长？是否支持数据并行存储？

为了解决单处理器的计算机编写的数据挖掘算法无法在并行计算机上自动以更快的速度运行的问题，为了充分发挥并行计算的优点，需要编写支持并行计算的算法。

3. 易操作性

易操作性是一个重要的因素。有的工具有图形化界面，引导用户半自动化地执行任务，有的使用脚本语言。有些工具还提供数据挖掘的 API，可以嵌入到像

C、Visual Basic 和 Power Builder 这样的编程语言中。

模式可以运用到已存在或新增加的数据上。有的工具有图形化的界面，有的允许通过使用 C 这样的程序语言或 SQL 中的规则集，把模式导入到程序或数据库中。

4. 数据存取能力

好的数据挖掘工具可以使用 SQL 语句直接从 DBMS 中读取数据。这样可以简化数据准备工作，并且可以充分利用数据库的优点（比如并行读取）。没有一种工具可以支持大量的 DBMS，但可以通过通用的接口连接大多数流行的 DBMS。Microsoft 的 ODBC 就是一个这样的接口。

5. 与其他产品的接口

有很多其他的工具可以帮助用户理解数据、理解结果。这些工具可以是传统的查询工具、可视化工具、OLAP 工具。数据挖掘工具是否能提供与这些工具集成的简易途径?

因为数据挖掘工具需要考虑的因素很多，很难按照原则给工具排一个优劣次序。最重要的还是用户的需要，根据特定的需求加以选择。数据挖掘工具可以给很多产业带来收益。国外的许多行业如通信、信用卡公司、银行和股票交易所、保险公司、广告公司和商店等已经大量利用数据挖掘工具来协助其业务活动。国内在这方面的应用还处于起步阶段，对数据挖掘技术和工具的研究人员以及开发商来说，我国是一个有巨大潜力的市场。

1.1.5 数据挖掘的需求与挑战

有效、流畅地处理大量数据的能力对数据挖掘提出了大量的需求与挑战。需求与挑战放在一起是因为在最近的数据挖掘研究与发展中，有些挑战已经得知并在一定程度被看成是需求，同时另外的还在研究阶段。下面从挖掘的方法论、用户的交互、性能、数据的形式、应用和社会的冲击几个方面来讨论需求与挑战。

1. 挖掘的方法论和用户交互的论点

这种论点将对挖掘知识的类型、在多个粒度上挖掘知识的能力、应用的知识领域、即席挖掘和知识的可视化等多方面产生影响。

(1) 在数据库中挖掘不同的知识。不同的用户对不同的知识感兴趣，数据挖掘应该能覆盖数据分析与知识发现任务的许多方面，包括分类或预测模型发现、数据总结、聚类、关联规则发现、序列模式发现、依赖关系或依赖模型发现、演变与偏离分析发现等。这些任务需要在相同的数据库上应用不同的方法，并需要