数据挖掘技术

吉根林[1,2] 孙志挥[2]
[1](南京师范大学计算机系, 南京 210097) [2](东南大学计算机系, 南京 210096)

摘 要 数据挖掘技术是当前数据库和人工智能领域研究的热点课题,为了使人们对该领域现状有个概略了解,在消化大量文献资料的基础上,首先对数据挖掘技术的国内外总体研究情况进行了概略介绍,包括数据挖掘技术的产生背景、应用领域、分类及主要挖掘技术;结合作者的研究工作,对关联规则的挖掘、分类规则的挖掘、离群数据的挖掘及聚类分析作了较详细的论述;介绍了关联规则挖掘的主要研究成果,同时指出了关联规则衡量标准的不足及其改进方法,提出了分类模式的准确度评估方法;最后,描述了数据挖掘技术在科学研究、金融投资、市场营销、保险业、制造业及通信网络管理等行业的应用情况,并对数据挖掘技术的应用前景作了展望.

关键词 数据挖掘 决策支持 关联规则 分类规则 KDD
中图法分类号: TP391 TP182 **文献标识码**: A **文章编号**: 1006-8961(2001)08-0715-07

Survey of the Data Mining Techniques

JI Gen-lin[1,2], SUN Zhi-hui[2]
[1](*Department of computer, Nanjing Normal University, Nanjing* 210097)
[2](*Department of computer, Southeast University, Nanjing* 210096)

Abstract Data mining is an emerging research field in database and artificial intelligence. In this paper, the data mining techniques are introduced broadly including its producing background, its application and its classification. The principal techniques used in the data mining are surveyed also, which include rule induction, decision tree, artificial neural network, genetic algorithm, fuzzy technique, rough set and visualization technique. Association rule mining, classification rule mining, outlier mining and clustering method are discussed in detail. The research achievements in association rule, the shortcomings of association rule measure standards and its improvement, the evaluation methods of classification rules are presented. Existing outlier mining approaches are introduced which include outlier mining approach based on statistics, distance-based outler mining approach, data detection method for deviation, rule-based outlier mining approach and multi-strategy method. Finally, the applications of data mining to science research, financial investment, market, insurance, manufacturing industry and communication network management are introduced. The application prospects of data mining are described.

Keywords Data mining, Decision support, Association rule, Classification rule, KDD

0 引 言

数据挖掘(Data Mining),也称数据库中的知识发现(KDD: Knowledge Discovery in Database),是指从大型数据库或数据仓库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信息,提取的知识一般可表示为概念(Concepts)、规则(Rules)、规律(Regularities)、模式(Patterns)等形式[1]. 大家知道,如今已可以用数据库管理系统来存储数据,还可用机器学习的方法来分析数据和挖掘大量数据背后的知识,而这两者的结合就促成了数

基金项目:国家自然科学基金项目(79970092)
收稿日期:2000-06-22;**改回日期**:2000-12-14

据挖掘技术的产生. 数据挖掘是一门交叉性学科,涉及到机器学习、模式识别、归纳推理、统计学、数据库、数据可视化、高性能计算等多个领域.

1989 年 8 月在美国底特律召开的第 11 届国际人工智能会议上首先出现 KDD 这个术语,随后引起了国际人工智能和数据库等领域专家的广泛关注. 1995 年在加拿大蒙特利尔召开了首届 KDD & Data Mining 国际学术会议,从此以后,KDD & Data Mining 国际学术会议每年召开一次. 经过十多年的努力,数据挖掘技术的研究已经取得了丰硕的成果,不少软件公司已研制出数据挖掘软件产品,并在北美、欧洲等国家得到应用[1]. 例如,IBM 公司开发的 QUEST 和 Intelligent Miner; Angoss Software 开发的基于规则和决策树的 Knowledge Seeker, Advanced Software Application 开发的基于人工神经网络的 DBProfile; 加拿大 Simon Fraser 大学开发的 DBMinner; SGI 公司开发的 MineSet 等. 在我国,数据挖掘技术的研究也引起了学术界的高度重视,已成为信息科学界的热点研究课题.

数据挖掘研究具有广泛的应用前景,因为数据挖掘产生的知识可以用于决策支持、信息管理、科学研究等许多领域. Parsaye 把决策支持空间从应用层次上分成数据空间(Data Space)、聚合空间(Aggregation Space)、影响空间(Influence Space)和变化空间(Variation Space)等 4 个子空间[2](见图 1).

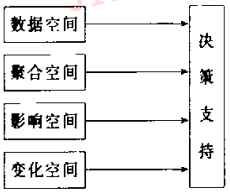

图1 决策支持空间

其中,数据空间是用于处理基于关键字的决策查询,其最典型的是联机事务处理(OLTP);而对数据空间中数据元素进行聚合运算(如 Sum, Average, Max, Min 等)所形成的空间就是聚合空间,它主要用于联机分析处理(OLAP);影响空间则用于处理逻辑性质的决策支持,比如回答"是什么因素影响公司的销售情况?"这样的问题,这些信息就是通过数据挖掘得到的;变化空间负责回答某种变化的过程和速度问题. 在上述 4 个空间中,数据挖掘处于影响空间中,从中可以看出数据挖掘在决策支持中所处的重要地位.

1 数据挖掘技术的分类

数据挖掘技术有根据发现知识的种类分类、根据挖掘的数据库种类分类、根据采用的技术分类等几种分类方法[3].

其中,根据发现知识的种类分类有关联规则挖掘、分类规则挖掘、特征规则挖掘、离群数据挖掘、聚类分析、数据总结、趋势分析、偏差分析、回归分析、序列模式分析等;根据挖掘的数据库种类分类有关系型、事务型、面向对象型、时间型、空间型、文本型、多媒体型、主动型和异构数据库等;根据采用的技术分类,最常用的数据挖掘技术有如下 7 种:

(1)规则归纳 即通过统计方法归纳、提取有价值的 if-then 规则,例如关联规则挖掘.

(2)决策树方法[4] 即用树形结构表示决策集合,这些决策集合是通过对数据集的分类来产生规则. 决策树方法是首先利用信息熵来寻找数据库中具有最大信息量的字段,从而建立决策树的一个结点,再根据字段的不同取值来建立树的分支;然后在每个分支子集中,重复建立树的下层结点和分支,即可建立决策树. 国际上最有影响的决策树方法是由 Quinlan 研制的 ID3 方法. 具体算法参见文献[4]. 其典型的应用是分类规则挖掘.

(3)人工神经网络[5] 这种方法主要是模拟人脑神经元结构,也是一种通过训练来学习的非线性预测模型. 它可以完成分类、聚类、特征规则等多种数据挖掘任务,同时它又以 MP 模型和 HEBB 学习规则为基础,来建立前馈式网络、反馈式网络、自组织网络 3 类神经网络模型.

(4)遗传算法[6] 这是一种模拟生物进化过程的算法,最早由 Holland 于 20 世纪 70 年代提出. 它是基于群体的、具有随机和定向搜索特征的迭代过程,这些过程有基因组合、交叉、变异和自然选择 4 种典型算子. 遗传算法作用于一个由问题的多个潜在解(个体)组成的群体上,并且群体中的每个个体都由一个编码表示,同时每个个体均需依据问题的目标函数而被赋予一个适应值. 另外,为了应用遗传算法,还需要把数据挖掘任务表达为一种搜索的问题,以便发挥遗传算法的优势搜索能力.

(5)模糊技术[7] 即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.这种模糊性是客观存在的,且系统的复杂性越高,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亦此亦彼性的,而李德毅教授在传统模糊理论和概率统计的基础上,提出了定性定量不确定性转换模型——云模型[8],并形成了云理论.云模型是用期望值、熵和超熵来表达定性概念,同时将概念的模糊性和随机性结合在一起,因而它为数据挖掘提供了一种概念和知识表达、定性定量转换、概念的综合和分解的新方法.

(6)粗(Rough)集方法[9]　它是 1982 年由波兰逻辑学家 Pawlak 提出的一种全新的数据分析方法,近年来在机器学习和 KDD 等领域获得了广泛的重视和应用.这种粗集方法是一种研究信息系统中不确定、不精确问题的有效手段,其基本原理是基于等价类的思想,而这种等价类中的元素在粗集中被视为不可区分的,其基本方法是首先用粗集近似的方法来将信息系统(关系)中的属性值进行离散化;然后对每一个属性划分等价类,再利用集合的等价关系来进行信息系统(关系)的约简;最后得到一个最小决策关系,从而便于获得规则.

(7)可视化技术[10]　即采用直观的图形方式来将信息模式、数据的关联或趋势呈现给决策者,这样决策者就可以通过可视化技术来交互地分析数据关系,而可视化技术主要包括数据、模型和过程 3 方面的可视化,其中,数据可视化主要有直方图、盒须图和散点图;模型可视化的具体方法则与数据挖掘采用的算法有关,例如,决策树算法采用树形表示;而过程可视化则采用数据流图来描述知识的发现过程.

上述数据挖掘技术虽各有各的特点和适用范围,但它们发现知识的种类不尽相同,其中规则归纳法一般适用于关联规则、特征规则、序列模式和离群数据的挖掘;决策树方法、遗传算法和粗集方法一般适用于分类模式的构造;而神经网络方法则可以用于实现分类、聚类、特征规则等多种数据挖掘;模糊技术通常被用来挖掘模糊关联、模糊分类和模糊聚类规则.

2　关联规则的挖掘

2.1　什么是关联规则

关联规则的挖掘[11]是数据挖掘领域中一个非常重要的研究课题,它是由 Agrawal 等人首先提出的.关联规则的挖掘问题可形式化描述如下:

设 $\boldsymbol{I}=\{i_1,i_2,\cdots,i_m\}$ 是由 m 个不同的项目组成的集合,给定一个事务数据库 D,其中的每一个事务 $\boldsymbol{T}$ 是 $\boldsymbol{I}$ 中一组项目的集合,即 $\boldsymbol{T}\subseteq\boldsymbol{I}$,$\boldsymbol{T}$ 有唯一的标识符 TID.一条关联规则就是一个形如 $\boldsymbol{X}\Rightarrow\boldsymbol{Y}$ 的蕴含式,其中,$\boldsymbol{X}\subseteq\boldsymbol{I}$,$\boldsymbol{Y}\subseteq\boldsymbol{I}$,$\boldsymbol{X}\cap\boldsymbol{Y}=\varnothing$.关联规则 $\boldsymbol{X}\Rightarrow\boldsymbol{Y}$ 成立的条件是:①它具有支持度 S,即事务数据库 $\boldsymbol{D}$ 中至少有 S%的事务包含 $\boldsymbol{X}\cup\boldsymbol{Y}$;②它具有置信度 C,即在事务数据库 D 所包含 $\boldsymbol{X}$ 的事务中,至少有 C%的事务同时也包含 $\boldsymbol{Y}$,关联规则的挖掘问题就是在事务数据库 D 中找出具有用户给定的最小支持度 $S_{\min}$和最小置信度 $C_{\min}$的关联规则.

挖掘关联规则可以分解为以下两个子问题:

①找出存在于事务数据库中的所有大项集.大项集是指支持度不小于用户给定的最小支持度的项集.

②利用大项集生成关联规则.对于每个大项集 $\boldsymbol{A}$,若 $\boldsymbol{a}\subset\boldsymbol{A}$,$\boldsymbol{a}\neq\varnothing$,且 Support($\boldsymbol{A}$)/Support($\boldsymbol{a}$)$\geqslant C_{\min}$,则 $\boldsymbol{a}\Rightarrow\boldsymbol{A}-\boldsymbol{a}$.这里,Support($\boldsymbol{A}$)、Support($\boldsymbol{a}$)分别表示 $\boldsymbol{A}$ 和 $\boldsymbol{a}$ 的支持度.

第②个子问题比较容易,其生成算法可参见文献[11].目前大多数研究均集中在第一个子问题上,因为这个问题的主要挑战性在于数据量巨大,所以算法的效率是关键.

3.2　关联规则的研究现状及研究领域

如今,关联规则的挖掘已经取得了令人瞩目的成果,到目前为止,主要研究工作有:

(1)多循环方式的挖掘算法[12]

多循环方式的挖掘算法是关联规则挖掘的基本方法.此类算法包括 Agrawal 等人提出的 AIS;Apriori 和 AprioriHybrid,Park 等人提出的 DHP 算法和分割算法 Partition 以及 Toivonen 提出的抽样算法 Sampling 等等.其中,Apriori 算法的基本思想是重复扫描数据库,并在第 K 次扫描时产生出长度为 K 的大项集 $\boldsymbol{L}_K$,而在第 $K+1$ 次扫描时,只考虑由 $\boldsymbol{L}_K$ 中的 K 项集产生长度为 $K+1$ 的备选集 $\boldsymbol{C}_{K+1}$;DHP 算法是使用 Hashing 技术来改进备选集 $\boldsymbol{C}_K$ 的产生过程;Partition 算法是将数据库进行分割,以减少挖掘过程中 I/O 操作次数;Sampling 算法则是首先对数据库进行抽样,然后对抽样数据库进行挖掘,从而提高了挖掘效率.国内研究人员还提出了一些 Apriori 算法的改进算法.

(2)并行挖掘算法

目前已经提出的有关并行挖掘关联规则的算法

有:Agrawal 等人提出的 CD(Count Distribution)算法、CaD(Candidate Distribution)算法、DD(Data Distribution)算法[13]和由 Park 等人提出的 PDM 算法,以及由 Chueng 等人提出的算法 DMA[14]算法和 FDM 算法,虽然这些算法均是基于分布式数据库的挖掘算法,但也适用于并行挖掘.

(3)增量式更新算法

关联规则的增量式更新问题主要有两种情况:①在给定的最小支持度和最小置信度条件下,当数据库添加了新记录后,如何生成数据库中的关联规则;②给定一个数据库,在最小支持度和最小置信度发生变化时,如何生成数据库中的关联规则. 文献[15],[16]已对关联规则更新问题进行了讨论,并提出了相应算法 FUP、IUA、PIUA 和 NEWIUA.

(4)基于约束条件的关联规则挖掘

基于约束条件的关联规则挖掘的主要目的就是发现更有趣、更实用、更特别的关联规则,文献[17]就研究了在提供布尔表达式约束情况下的关联规则发现问题.

(5)挖掘多值属性关联规则

关联规则可分为布尔型关联规则和多值属性关联规则,而多值属性又可分为数量关联规则和类别关联规则,其中数量关联规则是指同时包含布尔属性和连续属性的关联规则,如 Agrawal 等人扩展布尔属性的关联规则算法,就将其应用于数量关联规则的挖掘,并提出了基于支持度的部分 K 度完全方法;Fukuda 提出了等深度划分的实现方法[18];苑森森教授提出的在数量关联规则挖掘中的聚类方法 PKCCA[19]等. 目前提出的类别属性关联规则的挖掘算法,大多是将类别属性关联规则的挖掘问题转化为布尔型关联规则的挖掘问题[20],即将类别属性中的每一个类别当作一个属性.

2.3 关联规则衡量标准的不足

目前,生成关联规则的标准主要有如下两个,即支持度和置信度,但如果仅仅使用用户给定的最小支持度和置小置信度来生成关联规则,则往往会生成大量冗余的、虚假的和用户不感兴趣的关联规则.

下面用一个例子来说明这个问题. 如表 1 所示,设有 3 个项目数据集分别为 X,Y 和 Z,则可以发现关联规则 $X\Rightarrow Y$ 和 $X\Rightarrow Z$,其支持度与信任度见表 1.

但从表 1 中可看出,事实上 Z 与 X 之间并不相关,即 $X\Rightarrow Z$ 是一个虚假规则. 奇怪的是,虚假规则 $X\Rightarrow Z$ 的支持度和信任度却分别超过规则 $X\Rightarrow Y$ 的支持度和信任度,然而,还不可能找到合适的最低支持度和最低信任度,使得仅生成 $X\Rightarrow Y$,而不生成虚假规则 $X\Rightarrow Z$.

表 1 X、Y、Z 数据集及其相应的支持度、信任度

数据集			规则	支持度	信任度
X	Y	Z		(%)	(%)
1	1	0			
1	1	1	$X\Rightarrow Y$	25	37.5
1	0	1			
1	0	1			
0	0	1			
0	0	1	$X\Rightarrow Z$	50	75
0	0	1			
0	0	1			

这种问题已经引起了不少学者的注意,并提出在关联规则生成时要加限制条件,如将兴趣度这个标准加入到关联规则的定义之中.

3 分类规则的挖掘

3.1 分类的基本概念

分类是数据挖掘的一种非常重要的任务,它是在已有数据的基础上学会一个分类函数或构造一个分类模型(即通常所说的分类器),而且该函数或模型能够把数据库中的数据记录映射到给定类别中的某一个,从而可以应用于数据预测;若要构造分类模型,则需要有一个训练样本数据集作为输入,该训练样本数据集由一组数据库记录或元组构成,其一个具体的样本记录形式可以表示为 $(V_1,V_2,\cdots,V_n,C)$,其中,V_i 表示样本的属性值,C 表示类别.

3.2 分类模型的构造方法

分类模型的构造方法通常有统计方法(如贝叶斯方法)、机器学习方法(如决策树方法)、神经网络方法和等. 其中,基于统计方法的分类算法包括 Naive Bayes, K-nearest Neighbor, Kernel density, Linear discriminant, Quadratic discriminant, Logistic regression, Projection pursuit, Bayesian network 等算法;而基于机器学习的分类算法则包括 CART, C4.5, NewID, AC2, CAL5, CN2, Itrule 等;基于神经网络的分类算法包括 Backpropagation, Radial basis function, Kohonen 等;另外,基于粗集方法的分类方法国内外有关学者也提出了一些算法,如 RSBIDM[21]等.

其中,决策树方法、神经网络方法和粗集方法的

基本思想已在前面介绍，而贝叶斯方法的基本思想是：假定对研究对象已有一定的认识，那么即可先用先验概率分布来描述这种认识，然后用样本来修正已有的认识，得后验概率分布，最后通过后验概率分布来建立分类函数，其具体方法参见文献[22].

3.3　分类模式的准确度评估方法

由于分类模式正确率与训练集的记录数量、属性的数目及待测记录的分布等因素有关，且通常训练集越大，分类模式就越可靠，而属性数目越多，则生成分类模式的难度就越大，其需要的时间也越长，有时还会将分类器引入歧途，致使构造出不准确的分类模式，因此，如果可以通过常识确认某个属性与分类无关，则应将它从训练集中移走.

对产生的分类模式，可以用如下两种方法来进行准确度评估[23]：①保留方法(Holdout)，即将数据库中的一部分(通常是 2/3)作为训练集，而保留剩余的部分用作测试集，分类器是首先使用 2/3 的数据来构造分类模式，然后再使用该分类模式对测试集进行分类，其得出的正确率就是评估的正确率；②交叉纠错方法，即将数据集分成 K 个没有交叉数据的子集，且使所有子集的大小大致相同，这种分类器训练和测试共 K 次，且每一次，分类器使用其中 $(K-1)$ 个子集来作为训练集，然后在另一个子集上进行测试，最后把所有得到的正确率的平均值作为评估正确率.

4　聚类分析

聚类是数理统计中研究“物以类聚”的一种方法，它的任务是把一组个体按照相似性归成若干类，其目的是使得属于同一个类别数据之间的相似性尽可能大，而不同类别的数据之间的相似性尽可能小. 它与分类分析不同，聚类分析输入的是一组未分类的记录，并且这些记录应分成几类事先也不知道. 聚类分析就是首先通过分析数据库中的数据，合理地来划分记录，然后再确定每个记录所在类别. 另外，从技术上看，聚类分析可以采用统计方法、机器学习方法、人工神经网络方法、模糊技术来加以实现. 其中，在统计方法中，聚类算法一般分为基于概率的聚类算法和基于距离的聚类算法两种[22]，如欧氏距离等. 其中，基于概率的聚类算法在挖掘海量数据集合时效率非常低；而基于距离的聚类算法在数据挖掘领域应用则相当广泛，而且其基本思想是属于同一类别的个体之间的距离尽可能小，而不同类别上的个体间距离尽可能大.

5　离群数据挖掘

所谓离群数据，是指明显偏离其它数据，不满足数据一般模式或行为，即与存在的其他数据不一致的数据. 离群数据的挖掘是数据挖掘的重要内容，它包括离群数据的发现和离群数据的分析，其中离群数据的发现往往可以使人们发现一些真实的，但又出乎意料的知识；而离群数据的分析则可能比一般数据所包含的信息更有价值.

据研究，离群数据挖掘有着广阔的应用前景，例如，在数据分析时，错误数据的查找；金融、通信领域的欺诈分析与检测；网络安全管理中，网络入侵的检测；市场分析中，分析消费极高或极低的客户的行为；治疗过程中，异常反映的发现等. 目前这一领域正逐渐引起数据库、机器学习和统计学等方面学者的研究兴趣. 离群数据的发现主要有下列几种方法[24,25]：

(1)基于统计的离群数据发现方法

它是根据已知的数据分布模型，使用不一致性检验(discordance test)来确定离群数据，但它的应用需要事先知道数据集参数(如正态分布)、分布参数(如均值、标准差)和离群数据的个数，而且这种方法通常对数值型数据有效，而对高维、周期性数据、分类数据则较难进行挖掘.

(2)基于距离的离群数据发现方法

通过数据间距离的计算，即可求得离群数据. 设数据集 T，$T=t_1,t_2,\cdots,t_n$；o 为数据对象，如果数据集 T 中有 p 部分数据 S，远离于对象 o 及与之距离为 d 的邻域，则 o 即为基于距离 d 的离群数据，表示为 $DB(p,d)$. 如今基于距离的离群数据发现算法主要有以下几种：①Index-based 算法；②Nested-loop 算法；③Cell-based 算法.

(3)基于偏离的离群数据检测方法

这种方法是通过对各种形式的数据进行离群检测来发现离群数据. 但由于要事先知道数据的特性，以便确定相异函数；如相异函数的选取不合适，就得不到满意的结果，故较难在实际问题中使用.

(4)基于规则的分类数据离群发现方法

这种方法是从大量数据中产生离群数据的规则. 其主要步骤是：①首先根据属性值及其组合来构

成数据项集,且这种离群数据发现可以看作是树的搜索问题,其根结点是空条件项集,第 1 层结点是由长度为 1 的条件项集组成;然后计算某一条件项的支持度,以产生包含此结点,且长度为 2 的第 2 层结点;其他层次结点的产生方法依此类推;②根据多层最大离群支持度来求得离群规则.

(5)离群数据发现的多策略方法

这种方法首先对要挖掘的数据进行聚类,并将其分成具有不同特征的数据子集,这样目标范围小,特征更为明显,然后再从不同的数据子集中来产生规则.

6 数据挖掘应用

数据挖掘技术旨在发现大量数据中所隐藏的知识,以用来解决"数据丰富、知识贫乏"的问题.近年来随着数据库和网络技术的广泛应用,加上使用先进的自动数据生成和采集工具,人们所拥有的数据量急剧增加,为数据挖掘技术的应用创造了必要条件.目前国际上数据挖掘技术在科学研究、金融投资、市场营销、保险、医疗卫生、产品制造业、通信网络管理等行业[26,27]已得到应用;国内在数据挖掘方面也有成功的应用,例如宝钢已应用数据挖掘系统辅助生产决策,每年能节省近千万元资金.现在我国的研究人员正在加紧研制有关领域的数据挖掘工具,且数据挖掘技术的应用领域正不断扩大.

(1)科学研究　在信息量极为庞大的天文、气象、生物技术等领域中,由于所获得的大量实验和观测数据靠传统的数据分析工具已难以对付,因此对功能强大的智能化自动分析工具要求迫切,这种需求推动了 KDD 技术在科学研究领域的应用发展,并且已获得一些重要的应用成果,例如,美国加州理工学院喷气推进实验室与天文学家合作开发的 SKICAT 系统通过对几百万个天体进行分类,已帮助天文学家发现了 16 个新的类星体.

(2)金融投资　由于金融投资的风险很大,因此在进行投资决策时,需要对各种投资方向的有关数据进行分析,以选择最佳的投资方向,而数据挖掘则可以通过对已有数据进行处理,并利用学习得到的模式进行市场预测,例如,国内开发的指南针、神光、RMR 等智能股票分析系统,即可以对股票行情进行分析预测.目前作者正在利用数据挖掘技术研制一个智能股票分析系统.

(3)市场营销　主要用于商品的市场定位和消费者分析,以辅助制定市场策略;还可以用来分析购物模式,预测销售行情.例如,IBM 公司开发的 QUEST 和 Intelligent Miner 系统就可以挖掘顾客的购物行为模式.

(4)保险业　保险是一项风险业务,保险公司的一个重要工作就是进行风险评估.通过研究证明,可以利用数据挖掘来技术进行风险分析,在保险公司建立的保单及索赔信息数据库的基础上,寻找保单中风险较大的领域,从而得出一些实用的控制风险的规则,以指导保险公司的工作,例如,利用 SGI 公司的 MineSet 系统提供的分类器就可以预测投保人在将来的索赔概率.

(5)制造业可　制造业应用数据挖掘技术来进行零件故障诊断、资源优化、生产过程分析等,因为通过对生产数据进行分析,可发现容易产生质量问题的工序以及相关的故障因素等,例如,Acknosoft 公司开发的 CASSIOPEE 系统已用于诊断和预测在波音飞机制造过程中可能出现的问题.

(6)通信网络管理　在通信网络运行过程中,会产生一系列警告,虽然这些警告有的可以置之不理,而有的如果不及时采取措施,则会带来不可挽回的损失.由于警告产生的随机性很大,究竟哪些警告可以不予理睬,哪些警告必须迅速处理则往往很难判断,一般需要由人工根据经验来进行处理,因此效率不高,而数据挖掘则可以通过分析已有的警告信息的正确处理方法以及警告之间的前后关系,来得到警告之间的关联规则,这些有价值的信息可用于网络故障的定位检测和严重故障的预测,例如,芬兰 Helsinki 大学开发了一个基于通信网络中警报数据库的知识发现系统 TASA,将其用来寻找通信网络中警报序列规则,以便进行故障预测.

7 结　语

综上所述,数据挖掘涉及多种理论和技术问题,且它有着广泛的应用前景.最近的 Gartner 报告中就列举了今后 3～5 年对工业将产生重大影响的 5 项关键技术,而 KDD 技术就排列其中.数据挖掘在国外从理论研究到产品开发只用了 5～6 年时间,并且已经越来越多地用于大中型企业、商业、银行、保险业和电信业等部门,并表现出极强的发展潜力.数据挖掘这一新技术也必将在我国得到广泛的应用.

本文对近年来数据挖掘的主要研究内容进行了较全面的总结,还对数据挖掘中采用的主要技术手段作了分析,并介绍了数据挖掘技术的应用领域及作用,企盼对国内同行的研究有一定的参考价值.

参 考 文 献

1 Fayyad U M, Piatetsky-shapiro G, Smyth P. Advances in knowledge discovery and data mining. California: AAAI/MIT Press, 1996.

2 Kamran Parsage. Surveying decision support. Database Programming and Design. 1996,9(4):27～33.

3 胡侃,夏绍玮. 基于大型数据仓库的数据挖掘. 软件学报,1998,9(1):53～63.

4 陆汝钤. 人工智能. 北京:科学出版社,1996:823～844.

5 Lu Hongjun, Rudy Setiono, Liu Huan. Effective data mining using neural networks. IEEE Transactions on Knowledge and Date Engineering, 1996,8(6):957～961.

6 刘明吉,王秀峰,王治宝等. 一种基于遗传算法的知识挖掘算法. 计算机工程,2000,26(8):13～14.

7 何新贵. 数据采掘中的模糊技术. 计算机科学,1998,25(专刊):129～131.

8 李德毅,史雪梅,孟海军等. 隶属云和隶属云发生器. 计算机研究与发展,1995,42(8):32～41.

9 郭学军,陈晓云. 粗集方法在数据挖掘中的应用. 兰州大学学报(自然科学版),1999,35(增刊):276～279.

10 万家华,刘冰,江早. 知识发现中的可视化技术. 计算机科学,2000,27(增刊):131～134.

11 Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proceeding of the 20th international Conference on very large database, Santiago, Chile, Sept, 1994:487～499.

12 铁治欣,陈奇,俞瑞钊. 关联规则采掘综述. 计算机应用研究,2000,17(1):1～5.

13 Agrawal R, Shafer J C. Parallel mining of association rules. IEEE Transactions on knowledge and date engineering, 1996,8(6):962～969.

14 Cheung D W, Han Jiawei. Efficient mining of association rules in distributed database. IEEE Transactions on Knowledge and Data Engineering, 1996,8(6):910～921.

15 周海岩. 关联规则的开采与更新. 软件学报,1999,10(10):1078～1084.

16 冯玉才,冯剑琳. 关联规则的增量式更新算法. 软件学报,1998,9(4):301～306.

17 Srikant R, Agrawal R. Mining association rules with item constrains. In: Proceeding of the 3rd International Conference on KDD & Data Mining, Newport Beach, USA,1997:67～73.

18 Takeshi Fukuda, Yasuhiko Morimoto. Mining optimized association rules for numeric attributes. Journal of Computer and System Sciences, 1999,58(1):1～12.

19 苑森淼,程晓青. 数量关联规则发现中的聚类方法研究. 计算机学报,2000,23(8):866～871.

20 张朝晖,陆玉昌,张钹. 发掘多值属性的关联规则. 软件学报,1998,9(11):801～805.

21 淮晓永,熊范伦,赵星. 一种基于粗集理论的增量式分类规则知识挖掘方法. 南京大学学报(自然科学版,计算机专辑),2000,36(11):203～209.

22 方开泰. 实用多元统计分析. 上海:华东师范大学出版社,1992:189～193.

23 田金兰,李奔. 数据挖掘工具分类器. 计算机世界,1999.5.31.

24 史东辉,蔡庆生,倪志伟等. 基于规则的分类数据离群挖掘方法研究. 计算机研究与发展,2000,37(9):1094～1100.

25 史东辉,蔡庆生. 数据库中数据离群挖掘技术. 南京大学学报(自然科学版),2000,36(计算机专辑):82～86.

26 http://info.gte.com/～kdd

27 吉根林,帅克,孙志挥. 数据挖掘技术及其应用. 南京师大学报(自然科学版),2000,23(2):25～27.

吉根林　1964 年生,副教授,1989 年获南京航空航天大学计算机专业工学硕士学位,现为东南大学计算机系在职博士生.主要研究方向为数据库和数据挖掘技术.已发表论文 20 多篇.

孙志挥　现为东南大学计算机系教授,博士生导师.主要研究方向为数据库、数据挖掘和复杂信息系统集成技术.