数据挖掘工具 DMTools 的设计与实现

何耀东 常桂然 徐 茜 邵 华 李继晔
(东北大学软件中心，沈阳 110006)
E-mail:heyd@aries.synet.edu.cn

摘 要 介绍一个通用的数据挖掘工具 DMTools。它实现了基于数据库的知识发现(KDD)的主要过程：数据可视分析，数据预处理，数据库的知识发现，数据挖掘，模型解释及模型评估等。主要介绍了这个系统的体系结构和各模块的功能。使用本工具，可从各行业的历史业务数据库中挖掘出隐含的有价值的知识，用于决策支持。

关键词 数据库知识发现 数据挖掘 分类 聚类 关联规则

0 引 言

随着数据库技术的不断发展及数据库管理系统的广泛应用，数据库中存储的数据量急剧增大，从这些数据中挖掘出有用的信息(知识)以帮助人们作决策就成为一个急需解决的问题。为此，数据库中的知识发现(KDD)技术逐渐发展起来[1,2]。

KDD 处理过程可分为 9 个处理阶段，分别是：数据准备，数据选取，数据预处理，数据缩减，KDD 目标确定，挖掘算法确定，数据挖掘，模式解释及知识评价。如图 1 所示。数据选取的目的是确定发现任务的操作对象，即目标数据(target data)。数据预处理一般可能包括消除噪声、缺值数据处理、消除重复记录、完成数据类型转换(如把连续值数据转换为离散型的数据，以便于符号归纳，或是把离散型的转换为连续值型的，以便于神经网络归纳)等。数据缩减的主要目的是消减数据维数或降维，即从初始特征中找出真正有用的特征以减少数据挖掘时要考虑的特征或变量个数。

本文给出的通用数据挖掘工具 DMTools 实现了 KDD 的一些主要过程：数据可视分析，数据预处理，数据挖掘，模型解释及模型评估等。数据可视分析通过对数据进行可视分析以用于对数据进行缩减；数据预处理主要是对原始数据进行再加工，以便于从中提取知识；数据挖掘完成从数据中提取出用户所需要的知识；模型解释及模型评估对发现的知识进行评价并表达为用户易于理解的形式。

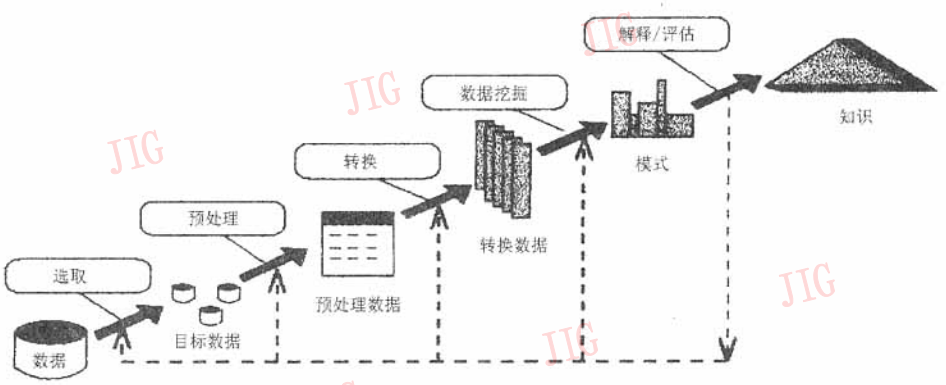


图 1 KDD 的一般过程

收稿日期：1999-10-27

1　系统的体系结构

本系统的体系结构见图 2。整个系统由数据可视分析、数据预处理、数据挖掘和模型解释及评估等 4 个模块组成。数据可视分析包括散点图、散点矩阵、特征选择和属性相似型分析等子模块。数据预处理包括离散化、采样等子模块；数据挖掘包括分类、聚类和关联规则发现等子模块；模型解释包括决策树可视化、聚类可视化和关联规则可视等；模型解释及评估包括分类精度检测和决策树可视、聚类结果评估和关联规则可视等子模块。

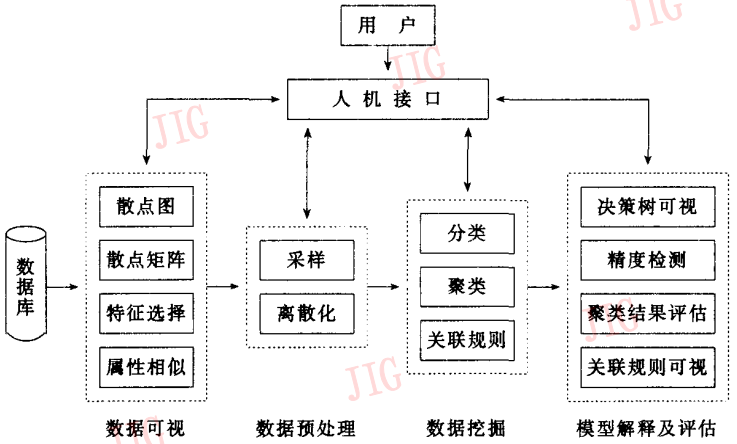


图 2　系统的体系结构

2　模块功能

2.1　聚类模块

聚类模块提供了多种聚类算法，可供用户选择使用，以找到最佳的聚类结果。这里提供的聚类算法主要有 k-均值、模糊 k-均值、ISODATA、系统聚类和 Autoclass 等 5 种聚类算法[3,4]。

2.2　关联规则发现模块

关联规则发现模块从交易数据集中发现隐含的关联规则。这里提供 Apriori[5]和 AprioriBest 两种关联规则发现算法。其中，AprioriBest 算法所需的执行时间比 Apriori 要少，但它所需的内存开销要大一些。由于所发现的关联规则可能很多，而其中有一些用户可能并不感兴趣，为此这里也提供规则过滤功能。过滤的方式有 3 种：基于最小支持度、最小可信度和 Lift 等的测度过滤；基于规定规则的两边所允许的项目集合的项目过滤；同时包含前两种方式的混合过滤。最后将这些规则保存在一个文本文件中，便于用户使用。

2.3　分类模块

分类模块对训练数据进行归纳，生成分类器，然后使用测试数据进行测试，最后把测试结果和分类器返回到模型可视模块。这里提供了多种分类算法，可供用户选择使用，以找到最佳的分类器。分类算法主要有 ID3、C4.5、Nbtree、IB、Navie bayes、HOODG 和 CN2 等 7 种算法。其中 ID3、C4.5 和 Nbtree 属于决策树，IB 属于基于实例的分类器，HOODG 属于决策图，Navie bayes 属于 bayes 分类器。

2.4　数据采样模块

应用采样技术处理源数据的原因是很显然的，采样技术可以大大减少数据挖掘所需要的时间。为了构造高精确度的数据模型，数据挖掘算法要求把原始数据集分为 4 个数据集。即：

(1) 训练集：用于输入到数据挖掘算法中构造模型。

(2) 控制集：用于控制数据挖掘算法避免过度训练(Over-training)。

(3) 测试集：用于评价特定模型的精确度。

(4) 检验集：用于评价最终模型的精确度。

一般的数据挖掘算法要求至少要有两个数据集，即训练集和测试集。为了能够正确地生成这些数据集需要正确的数据采样算法。这里给出的采样方式有自动采样和手工采样两种。

2.4.1 自动数据采样子模块

从原始数据集中提取两个数据集:训练集,测试集,并确保两个数据集中无交叉数据。自动采样提供两种方法:范围采样法,随机数采样法。

2.4.2 用户交互子模块

提供用户交互图形界面,允许用户手工采样(见图 3)。在图 3 中,打对号样本被分到训练集中,打叉号的样本被分到测试集中。

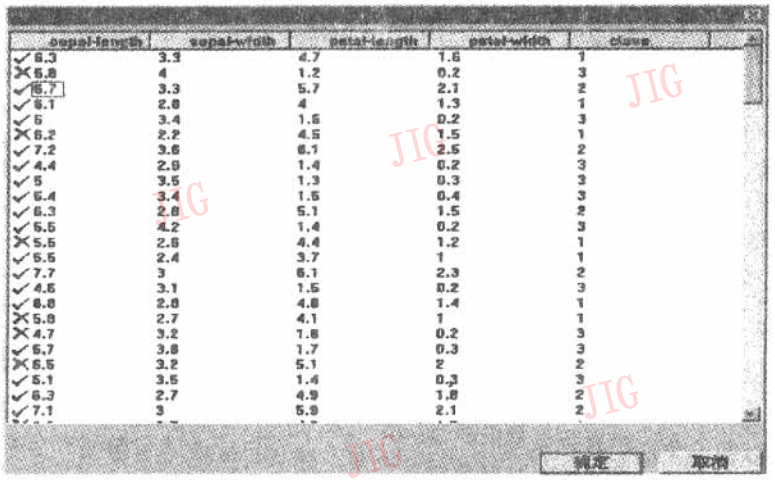


图 3 手工采样界面

2.5 模型解释及评估模块

可视化功能一方面允许用户以图形的方式理解各个抽象级别的数据,如:柱状图、饼图、曲线图等;另一方面允许用户利用图形的方式把数据挖掘的结果显示出来,如:决策树、关联规则等,使得用户可以直观地在图形界面下以不同的表示方式理解数据挖掘的结果,对挖掘的结果进行精度分析,充分地利用数据挖掘所获得的知识。

2.5.1 决策树可视模块

将挖掘结果以规则或树状的形式显示(如彩色图版 I 中的图 1),并能提供简单的用户交互,能让用户查询各个规则节点所对应的实例及实例个数,以及每个实例所对应的规则。

2.5.2 精度检测模块

给出利用训练集对所建立的分类器进行测试的结果,以助于对所建立的分类器进行评价。给出的信息主要包括训练集样本个数、测试集样本个数、测试正确样本个数和测试正确率。

2.5.3 聚类结果评估模块

通过结果评估、聚类结果、类分布、类间距离和结果可视等 5 个手段对聚类结果进行评估(如彩色图版 I 中的图 2)。结果评估给出数据集被聚的类别个数和聚类误差(在类别总数相等的情况下,聚类误差越小,聚类质量越高);聚类结果以表格的形式给出聚类结果;类分布以饼状图的形式给出属于每个类别的样本个数(若属于某个类别的样本个数很少,则这个类别应被合并);类间距离通过画出每个类的凝聚点以可视化的形式给出类别之间的距离(若几个类的凝聚点靠的很近,则说明这几个类应被合并为一个类)。结果可视通过画出这个数据集的散点图且属于同一类的样本用同一种颜色表示来直接可视地对聚类结果进行评价。

2.5.4 关联规则可视模块

这里以表格的形式将所发现的关联规则显示出来(如图 4),其中包括关联规则、每条关联规则的支持度、可信度和 Lift 等信息。这些信息既是规则过滤的基础,又是衡量所得规则的可靠性的依据。

2.6 数据可视分析模块

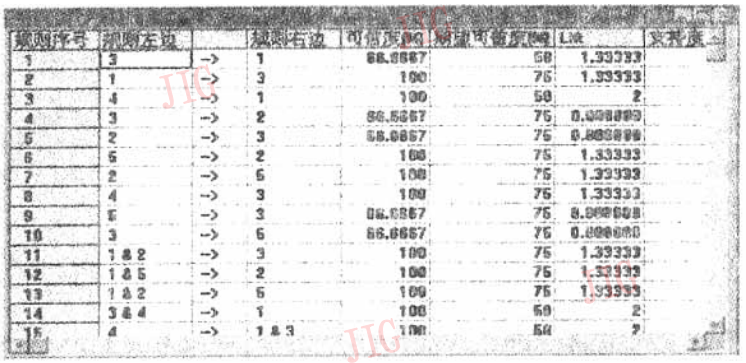
在数据准备阶段,用户可能要使用各类可视化技术来显示有关数据,以期对数据有一个初步的理解,从而为更好地选取数据打下基础。

2.6.1 特性选择模块

利用 K-W 检验法,对数据进行特征选择,并以图表和表格的形式进行可视分析。通过特征选择,可以淘汰特征分类能力弱的特征,而选择分类能力强的特征,以用于数据缩减,这样既可以提高模型的准确度,又可以缩短建模的时间以提高效率。

2.6.2 属性相似性子模块

分析属性的相似性,得出相似系数,并以图形和图表的形式显示。通过属性相似性,我们可以将相似



规则序号	规则左边		规则右边	可信度(%)	期望可信度阈值	Lift	支持度
1	3	->	1	66.6667	50	1.33333	
2	1	->	3	100	75	1.33333	
3	4	->	1	100	50	2	
4	3	->	2	66.6667	75	0.888889	
5	2	->	3	66.6667	75	0.888889	
6	5	->	2	100	75	1.33333	
7	2	->	5	100	75	1.33333	
8	4	->	3	100	75	1.33333	
9	5	->	3	66.6667	75	0.888889	
10	3	->	5	66.6667	75	0.888889	
11	1 & 2	->	3	100	75	1.33333	
12	1 & 5	->	2	100	75	1.33333	
13	1 & 2	->	5	100	75	1.33333	
14	3 & 4	->	1	100	50	2	
15	4	->	1 & 3	100	50	2	

图 4　关联规则可视

系数较大的两个属性合为一个属性或者只选取其中的一个属性。

2.6.3　散点图数据描述模块

利用主成分分析法和基于象素的可视技术画出数据集的散点图,通过散点图可以看出所有样本点在空间的分布情况,对这个数据集有个总体印象。另外散点图也可帮助确定聚类所要聚的类数以及发现含噪的数据。

2.6.4　散点矩阵表示法

散点矩阵法[6]揭示了属性之间的相关性。对于 k 维的数据集,将屏幕分为 $k\times k$ 个窗口,x,y 坐标分别对应 k 个属性,在每个窗口分别画出只考虑相应两个属性属性值的散点图(如图 5),从中可以看出一个属性的变化对另一个属性的影响情况。

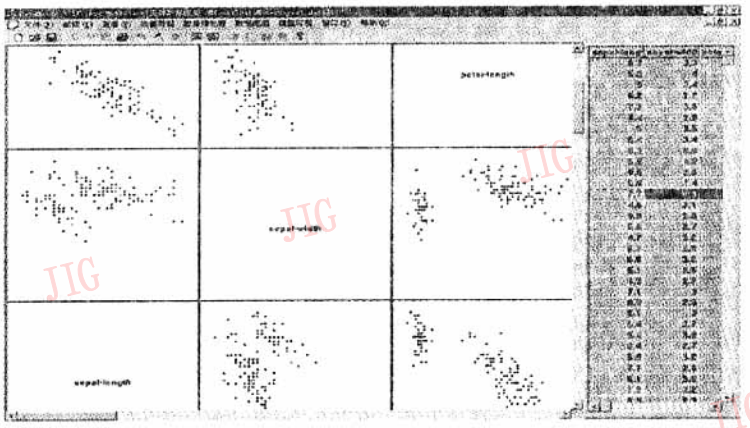


图 5　散点矩阵

3　结论及展望

本文主要介绍了通用数据挖掘工具 DMTools 的体系结构和各模块的功能。本系统是一个通用数据挖掘工具的原型系统,在一些方面还需要进一步完善和补充。例如在数据预处理模块,可以补充消除噪声、缺值数据处理等功能;在数据挖掘模块,可以补充序贯模式、神经网络、回归分析等模型。另外还应补充数据缩减功能,以达到降维的目的。通过进一步完善和补充,最终可形成一个较完备的通用数据挖掘工具。

参 考 文 献

1　Fayyad U M, Piatetsky-Shapiro G, Smyth, P. From data mining to knowledge discovery: An overview. In: Advances in Knowledge Discovery and Data Mining, Fayyd U M, Piatetsky-Shapiro G(eds), 1～35.

2　Brachman R J, Fnand T. The process of knowledge discovery in databases: A human-centered approach. In:Advances in Knowledge Discovery and Data Mining, Fayyad U M, Piatetsky-Shapiro G(eds), 37～58.

3　王碧泉,陈祖荫. 模式识别. 地震出版社. 1989.

4　Cheeseman P, Stutz J. Bayesian Classification(AutoClass):Theory and Results. In:Advances in Knowledge Discovery and Data Mining, Fayyad U M, Piatetsky-Shapiro G(eds), 153～180.

5　Agrawal R, Srikant R. Fast algorithms for mining association

rules. IBM Research Report RJ9839, June 1994, IBM Almaden Research Center, San Jose, Calif.

6 Ohavi R, Sommerfield D, Dougherty J. Data mining using MLC ++, a Machine Learning Library in C++. International Journal of Artificial Intelligence Tools, 1997, 6(4): 537～566.

Design and Realization for the Tools of Data Mining (DM-Tools)

He Yaodong, Chang Guiran, Xu Qian, Shao Hua and Li Jiye

(The Software Center of the Northeast University, Shenyang 110006)

Abstract The article introduces the common tools of Data Mining (DM-Tools), which realized the major process of Knowledge Discovery from Database (KDD). The process comprises visual data analysis, data preprocess, data mining, model explain and model evaluation, etc. System structure and module functions of the DM-Tools are discussed. Based on the DM-Tools, hidden and useful knowledge from long-term accumulation, operation database can be mined to support decision-making for every profession

Keywords KDD, Data mining, Classification, Cluster, Association rule

(下转第 956 页)