数据挖掘原型系统中分类挖掘模块设计与实现

吴湘洲 田盛丰 TP3 A

摘 要：介绍了通用数据挖掘原型系统 GenMiner 中分类挖掘模块设计与实现。GenMiner 系统中分类挖掘采用耗时短，分类效率高，较为成熟的决策树 C4.5 方法。文中说明了分类挖掘模块采用的决策树 C4.5 方法，及其在 GenMiner 系统设计与实现。

关键词：数据挖掘 GenMiner 分类 决策树 C4.5

Design and Implementation of the Classification Module in GenMiner

WU Xianzhou TIAN Shengfeng

(Northern Jiaotong University,Beijing,100044)

Abstract: The design and implementation of the classification module in GenMiner,which is a general datamining prototype system developed by us, is proposed in this paper. The classification module in GenMiner uses the decision tree C4.5 whose time-cost is small and classification is very efficient,and which has been developed very well.The paper addresses the method of decision tree C4.5 which is used in the classification module,and its design and implementation in GenMiner.

Key words Datamining; GenMiner; Classification; Decision tree; C4.5

1 引言

近几年来，随着数据库技术的迅速发展和管理系统的广泛应用，人们积累的数据越来越多。数据的背后隐藏着许多重要信息，人们希望能够对其进行更高层次的分析，以便更好地利用这些数据。目前的数据库系统可以高效地实现数据的录入、修改、统计、查询等功能，但无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势。缺乏挖掘数据背后隐藏的知识的手段，导致了“数据爆炸但知识贫乏”的现象。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。

KDD 是从数据集中识别出有效的、新颖的、潜在的、有用的以及最终可理解模式的高级处理过程。数据挖掘是 KDD 的核心部分。数据挖掘（Data Mining）是近几年人工智能和数据数据库技术研究的热点。数据挖掘就是从大量的数据中抽取以前未知并具有潜在可用的模式。数据挖掘技术表现出的广阔的应用前景吸引了众多的研究人员和商业机构，一批数据挖掘系统被开发出来，并在商业、经济、金融、管理等领域都取得了应用性成果。

我们开发的通用数据挖掘系统 GenMiner 是基于数据库的。本系统主要由五大模块组成，包括数据接口、数据离散化、关联规则挖掘、分类挖掘及结果可视化。本文重点介绍分类挖掘模块的设计和实现。

吴湘洲 北方交大人工智能实验室 硕士研究生 100044 北京市
田盛丰 北方交大人工智能实验室 教 授 100044 北京市

2 数据挖掘系统

数据挖掘系统的输入是数据库（或数据仓库）的数据、信息分析员的指导以及存储在挖掘系统知识库中的知识和规则。选择的数据在各挖掘模块中处理，生成辅助模式和关系。然后进行评价，通过与分析员交互以期发现令人感兴趣的模式。有些发现还要加入知识库中，以便后继的抽取并进行评价。数据挖掘系统由以下构件联合组成：

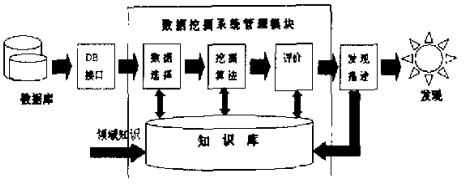

图 1 数据挖掘逻辑模型

3 GenMiner 数据挖掘系统简介

我们开发的通用数据挖掘系统 GenMiner 是基于数据库上的。系统主要由 5 大模块组成，包括数据接口、数据离散化、关联规则挖掘、分类挖掘及结果可视化。

3.1 数据接口模块

提供本原型系统和数据库访问接口。

3.2 数据预处理

本模块的目的是对原始数据进行处理，生成数据挖掘工具可利用的数据。

3.3 关联规则挖掘模块

包括若干种关联分析模型的工具。

3.4 分类规则挖掘模块

分类分析模型建造工具的实现，决策树 C4.5 法。本文重点讨论此部分。

3.5 结果可视化模块

对挖掘结果进行解释评估，给出各种直观的图形化显示方法。

4 分类挖掘模块的设计和实现。

4.1 分类挖掘及 C4.5 算法

分类属于“带监督”的机器学习，它的目的是学会一个分类函数或分类模型，即我们通常所说的分类器。分类器能够把数据库中的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。分类器的方法有决策树方法(Decision Tree)、神经网络方法(BP 算法)、遗传算法(Genetic Arithmetic)等等。其中决策树方法是很重要也是到目前为止发展最为成熟的一种概念学习方法。构造决策树分类器耗时短，分类效率高。本系统分类模块中实现了决策树 C4.5 的分类挖掘。

决策树方法起源于概念学习系统。J.R.Quinlan 提出 ID3 算法，J.C.Schlimmer 和 D.Fishgf 1986 年以及 P.E.Utgoff 于 1988 年又进一步提出 ID4 和 ID5R 算法，最后 J.R.Quinlan 于 1993 年提出了能处理连续属性的 C4.5，决策树方法发展得比较成熟。GenMiner 系统中的分类挖掘是采用决策树 C4.5 算法。

决策树是由一个根结点，若干的叶子结点和若干的非叶子结点构成的二叉树或多叉树。根结点对应于学习任务，每个叶节点都一个分类名，即包含一个概念。树的非叶子节点一般表示为一个逻辑判断。

要构造一个决策树，需要一个类别已知的训练集。设给定的训练集合 T，T 的元素由特征向量及其结果表示，分类对象的属性为 a1,a1...an，对于每一个属性 ai,其值域为 di。全部分类结果构成的集合 Class 为{C1，C2，C3...，Cm}.这样，T 的一个元素就可以表示成<X，C>的形式，其中 X={a1,a2,...，an}，ai 对应与该实例第 i 个属性的取值，C∈Class,为实例的分类结果。记$_{v}$(X,ai)为特征向量 X 的 ai 属性的取值。以下是决策树的构造算法 ID3：

(1) 如果 T 中所有的分类结果均为 Ci，则返回 Ci；

(2) 从属性表中选信息增益(或信息增益比)最大的属性 a 作为测试属性；

(3) 假设|di|= k，则根据 a 取值不同，将 T 划分为 T1，T2，...，Tk，其中 Ti={<X，C>|<X，C>∈T，且 V(X，A)为属性 A 的第 i 个值}；

(4) 标记已测试过的属性；

(5) 对每一个 i(1 ≤ i ≤ k)，用 Ti 和新的属性表调用 ID3 生成 Ti 的决策树 DTi；

(6) 返回以 DT1，DT2，...，DTk 为子树的决策树。

此处信息增益和信息增益比是熵的函数，熵表示平均信息量。通过此函数评价使用各属性进行分类获得的信息量，选择一个获得信息最大的用于分类，提高分类效率。

C4.5 算法是 ID3 算法的改进，C4.5 算法在 ID3 算法的原型基础之上介绍了大量的扩展。C4.5 算法在以下几个方面提高了 ID3 算法的性能：

（1）可以处理连续值属性；

（2）可以处理缺省值；

（3）引入修改方法；

（4）可以生成分类规则；

（5）分类算法的性能评估主要有以下标准：

（6）预测正确率，速度，鲁棒性等等。

4.2 GenMiner 分类模块的设计与实现

分类过程主要为分类的数据预处理，参数和类的选择，决策树的构造和修剪，分析和评估，生成分类规则。GenMiner 中的分类挖掘采用决策树 C4.5 挖掘算法。系统的分类挖掘结果是生成 C4.5 决策树，比较直观，易于理解。分类挖掘可以用不同于训练数据的测试数据对挖掘的结果进行分析和评估，而且可以生成易于理解的分类规则。

4.2.1 分类的数据预处理

现在的数据库通常数据量十分巨大，一条记录的属性可能是很多的，计算量非常大，而且也可能产生不必要的复杂分类规则。将原始数据进行一些预处理，会有利于提高分类正确率以及速度。用于本系统分析的数据是铁道部的货运数据，所以根据数据特点，对其进行必要的预处理，有利于后面的工作。

本系统中分类挖掘的预处理主要是基于冲突分析的特征提取和属性的离散化。特征提取是在众多属性中找出其取值能决定元组类标志的那些属性的过程，这样做能减少分类算法的执行时间，对于基于决策树的分类器，属性数量的减少,意味着更少的计算和更少的比较，由于与分类函数无关的属性预先被剔除了，分类器产生的规则可望更简明。属性的离散化是将属性的范围分为区间形式，区间标志取代实际的数据值，减少属性值数，属性离散化后，分类过程只需处理相对原始数据较少的属性值，有利于提高决策树分类的速度。

4.2.2 参数的选择和类的选择

在这一过程中，用户可以根据不同的要求来设置分类挖掘各项参数，得到较为满意的结果。这些参数设置包括是否生成规则集、是否采用子集法、是否模糊界限、是

否采用增量比准则、是否用测试数据评估、是否采用窗口法、生成树的数目、详细程度、信任度和最小例子数。用户对分类挖掘的类别可以进行选择设定。通过此交互过程，使得分类能满足用户需求，提高分类的正确率。这些参数提供出下一步决策树的构造和修剪。

4.2.3 决策树的构造和修剪

根据决策树归纳算法C4.5，从训练数据中学习构造决策树分类器，训练数据是由一系列已知分类的数据组成的集合。当树构造好之后，要对建立在噪音数据之上的分支进行修剪，最后得到一棵决策树。

GenMiner系统中采用后剪枝方法，是从“完全生长”的树剪去分枝。对于树中每个非树叶节点，计算该节点上的子树被剪枝可能出现的期望错误率。然后，使用每个分枝的错误率，结合沿每个分枝观察的权重评估，计算不对该节点剪枝的期望错误率。如果剪去该节点导致较高的期望错误率，则保留该子树；否则剪去该子树。最后得到具有最小期望错误率的判定树。

4.2.4 分析和评估

构造出分类器以后，必须对其进行分析和评估，知道它对未来数据进行分类的错误率是很有用的，对分类器的错误率的评估方法是假定待预测记录和训练集取自同样的样本分布。

在本系统中，采用了保持方法和K-折交叉确认方法。保持方法，将记录集中的2/3作为训练集，保留剩余的部分作测试集，生成器使用2/3的数据来构造分类器，然后使用这个分类器对测试集进行分类，得到的错误率就是评估错误率。这种方法速度很快。K-折交叉确认的方法：数据集被分成k个没有交叉数据的子集，所有子集的大小基本相同。生成器的训练和测试共用k次，每一次生成器使用除去一个子集的所有剩余数据作为训练集，然后在被除去的子集上进行测试，把所有得到的错误率的平均值作为评估错误率，交叉纠错法可以被重复多次，对于一个t次k分的交叉纠错法，k*t个分类器被构造并评估，这意味着交叉纠错法的时间是分类器构造时间的k*t倍，增加重复次数会导致运行时间的增长和错误评估的改善。

4.2.5 生成分类规则

此过程是将决策树转化为比较直观的规则形式，让用户能更好地理解分类结果。分类规则是用if-then形式表示，每一条规则都是一条从根到叶节点的路径，叶节点表示具体的结论，而叶节点以上的结点及其边表示的相应的条件的条件取值。如if(fee>200&fee<6000) then 类别=第4类。

本系统是在Windows2000操作系统下采用Visual C++6.0开发的，使用的数据库是Oracle。本系统各种大小类型的货运数据集上运行，得到比较满意的结果。

5 结束语

通用数据挖掘系统GenMiner是一个通用的知识发现工具。其中的分类挖掘模块是采用决策树C4.5算法，C4.5算法是一种适用范围比较广泛、效率较高的决策树算法。通过数据预处理，参数和类选定，构造和修剪决策树，进行分析和评估，生成分类规则等步骤后，完成分类挖掘，对各种大小类型的货运数据集进行测试，能得到比较满意的挖掘结果。

6 参考文献

1.G. Piatetsky-Shapiro, Knowledge Discovery in Databases, AAAI/MIT Press,1991.

2.Jiawei Han,Micheline Kambr. Data Mining—Concepts and Techniques, Morgan Kaufmann Pulishers,2000.

3.Alex Berson, Stephen J. Smith. Data Warehousing,Data Mining, &OLAP,McGraw-Hill Book Co,1999

4.JRQuinlan. Learning with continuous classes, In Procedings AI'92 (Adams Sterling, Eds),343-348,Singapore: World Scientific,1992

5.Quinlan, J. R.. C4.5: Programs for Machine Learning, 1993

6.C. Apte and SM Weiss. Data mining with decision trees and decision rules, Future Generation Computer Systems November 1997:197-210.

（收稿日期：2001-10-30）

3月份热门病毒预报

求职信变种病毒（Klez.e、Klez.f、Klez.g），危害★★★★★，近日，瑞星公司截获到“求职信”病毒的最新变种，并发现此病毒正通过电子邮件大规模扩散。它在原病毒的基础上增加了更多的破坏手段，当病毒驻留系统后会强迫关闭用户正在进行的其它正常操作，并会删除一些有用文件，所以较之以往的求职信病毒它具有更大的破坏力，此病毒多以以电子邮件的形式传播，其信件主题多为“Hi,Hello，Re:，how are you”的英文短句，而病毒邮件附件的扩展名则以：txt、htm、html、wab、doc、xls、jpg、cpp、cpas、mpg、mpeg、bak、mp3为主，因此，当用户收到具有以上特征的英文信件时，需谨慎处理！

发作病毒：Hack.qqph此病毒主要破解oicq密码，泄漏用户信息。**Hack.mobile.smsdos**此病毒感染win98/win2000。**SWF/LFM-926**病毒依赖于“Standalone Player”的FLASH浏览器。（瑞星公司提供）