部分数据缺失环境下的知识发现方法*

王清毅, 蔡 智, 邹 翔, 蔡庆生

(中国科学技术大学 计算机科学技术系,安徽 合肥 230027)

E-mail: qywang@ustc.edu.cn

http://www.ustc.edu.cn

摘要: 介绍了目前的不完全数据环境下的知识发现研究工作,分两个部分提出了一个不完全数据库中的知识发现方法.首先具体讨论了如何猜测丢失的数据,给出了基于距离的关联规则的定义及挖掘方法.然后在此基础上详细描述了一个不完全数据库中的知识发现算法,分析了算法的复杂度,并给出了相应的实验结果.最后,将所提方法与其他相关方法进行了比较.

关键词: 丢失数据;不完全数据库;知识发现;聚类;基于距离的关联规则

中图法分类号: TP18 **文献标识码**: A

目前的知识发现研究大都认为所采用的数据是完全的,因而所提出的方法和所开发出的系统仅适用于数据完全的环境.例如,目前的关联规则发现算法[1~3]在数据库中含有丢失的数据时就得不到令人满意的结果[4],这是因为这一算法是面向数据完全的数据库开发的.然而,我们在研究中发现,现实世界的数据库(例如,商业数据库和医院数据库)中的数据很少是完全的:丢失的数据、观测不到的数据、隐藏的数据、录入过程中发生错误的数据等等是现实世界数据库的一个常见特征.为了能在现实世界中有效地应用知识发现的方法和系统,就必须面对数据不完全的挑战.

1 相关的研究

人工智能领域的研究者已经提出了一些从含有丢失数据的训练例子集中生成决策规则的方法,其中最简单的就是去除带有丢失数据的例子,或者用最常出现的数据值代替丢失的数据.Kononenko 等人[5]采用贝叶斯方法确定丢失值在一个范围内取值的概率分布,要么取最可能的值,要么取将对象再分解成子对象,每个子对象根据各自确定的概率取一个加权的值.Quinlan 等人[6]建议,基于其他已知属性的值和分类信息来预测一个数据的丢失值.也有不少研究者研究了基于粗糙集理论的不完全系统中的知识发现方法,例如,Chmielewski 采用的方法是将不完全系统转变成完全系统,不完全系统中数据不完全的每个对象都由完全系统中的一组可能的子对象来描述,这种方法可以生成所有的确定规则.

Lakshminarayan 等人在已有的机器学习系统 AutoClass 的基础上开发了一个可处理丢失数据的知识发现系统[7].系统的任务之一首先是要把数据集分成多个类,并且对每个分类确定每个属性取值的概率分布.对一个测试数据集来说,不是直接预测数据的属性值,而是对每项数据 x,系统

* 收稿日期: 2000-01-13; 修改日期: 2000-06-26

基金项目: 国家自然科学基金资助项目(69975016)

作者简介: 王清毅(1962-),男,安徽黄山人,博士,讲师,主要研究领域为机器学习,知识发现;蔡智(1974-),女,湖北汉川人,博士生,主要研究领域为机器学习,知识发现;邹翔(1977-),男,安徽马鞍山人,硕士生,主要研究领域为机器学习,知识发现;蔡庆生(1938-),男,湖北汉川人,教授,博士生导师,主要研究领域为人工智能,机器学习,知识发现.

提供一个类的集合 C 上的概率分布,即 $p(\text{calss}(x)=C)$. 对于一项数据来说,已知它的类属,就可以利用该类的条件概率来推测出丢失的属性值. 例如,假定系统在经过学习以后对某项测试数据进行分类,将其归为 $C1$ 类的概率为 0.8,归为 $C2$ 类的概率为 0.2,如果该项数据的一个离散型属性的值丢失,系统就可以从 $C1$ 中取出最可能的值来替代被丢失的值.

Ragel 在文献[4]中提出了一种在含有丢失数据的数据库中发现关联规则的方法. 这一方法是将原先含有丢失数据的数据库依据不同的项集划分成若干个有效数据库,同时,对原先的支持度和可信度的定义进行了修改,在此基础上进行关联规则的挖掘. Ragel 等人在现实世界的一个含有丢失值的医院数据库中进行了实验,并报告了实验结果.

2 方法讨论

2.1 猜测丢失的数据

这里,我们利用了主成分分析法[3]. 主要思想是,通过主成分和已给出的数据来猜测丢失的数据. 具体的猜测过程主要由以下 3 个步骤组成:(1) 求出原始数据矩阵的协方差矩阵,确定该协方差矩阵的特征值和特征向量;(2) 选择出 K 个特征向量,即线性关联规则;(3) 猜测出被丢失的值.

2.1.1 确定原始数据矩阵的协方差矩阵、特征值和特征向量

我们以一个顾客——商品矩阵 D 为例来加以讨论. 设 D 中有 T 项交易,M 个商品项,目标是要发现协方差矩阵 C、特征值 eigenvalue 和特征向量 eigenvector. 下面是确定原始数据矩阵的协方差矩阵、特征值和特征向量的算法,其中 colavg 代表协方差矩阵中每一列的列平均值.

(1) For $i=1$ to M

  For $j=1$ to M

    初始化 colavg[j]=0.0;

    初始化 $C[i][j]$=0.0;

(2) For $i=1$ to T D 中每一交易 t_i

  For $j=1$ to M

    colavg[j]=colavg[j]+$t_i[j]$;

    If $t_i>0$ and $t_j>0$

      $C[i][j]=C[i][j]+t_i[i]*t_j[j]$;

(3) For $i=1$ to M

  colavg[i]=colavg[i]/T;

(4) For $i=1$ to M

  For $j=1$ to M

    $C[i][j]=C[i][j]-T*\text{colavg}[i]*\text{colavg}[j]$;

(5) 求出协方差矩阵 C 的特征值 eigenvalue 和特征向量 eigenvector,从大到小排序特征值和相应的特征向量.

2.1.2 选择 K 个特征向量

实际上,主成份在坐标轴上标出的方向就是原顾客-商品数据矩阵的协方差矩阵的特征向量. 这样,只要求出了这个协方差矩阵的特征向量,我们就可以得到相应的 M 个主成分. 但我们不必取所有的主成分来考虑. 我们选用下列方法来确定所需的特征向量,即主成分的个数 k:

$$\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{j=1}^{M}\lambda_j}\geqslant 85\%, \tag{2.1}$$

其中 λ_i 为上述协方差矩阵的特征值.

2.1.3 猜测丢失的值

在确定出 K 个主成分之后,就可以再根据已知的数据,猜测出丢失的数据.设 V 是由 K 个特征向量形成的 $M\times K$ 的矩阵,X_h 是含有 H 个被丢失值的向量,X 是已知的部分值,则猜测过程可以描述如下:

(1) 从 X_h 取出已知的部分值,形成向量 X;

(2) 从 V 中去除若干行,这些行数与 H 个被丢失的值的维数相等,形成新的矩阵 $V1$;

(3) 求 $V1$ 的逆矩阵 $V1^{-1}$;

If $H+K=M$

求方阵 V 的逆矩阵;

If $H+K<M$

求矩阵 V 的 pseudo-inverse 逆矩阵;

If $H+K>M$

去除 $H+K-M$ 个主成分,按 $H+K=M$ 的情况处理;

(4) $V\times V1^{-1}\times X$;

(5) 从以上乘积结果中得到猜测出的值.

2.1.4 方法的准确度度量

我们是通过所求得的主成分和已知的数据来猜测丢失值的,这样就存在一个猜测准确度的问题.我们利用以下方法来评价此方法的准确度:假设一个数据 X_{ij} 被丢失,对它的猜测值为 $X1_{ij}$,则此单个数据的猜测误差为 $X_{ij}-X1_{ij}$,而对整个 $N\times M$ 数据集中的丢失值,其总猜测误差为

$$RMS=\sqrt{\frac{1}{N\times M}\sum_{i=1}^{N}\sum_{j=1}^{M}(X1_{ij}-X_{ij})^2}. \tag{2.2}$$

就我们所知,目前的关联规则发现研究很少涉及到对所挖掘的规则的准确度进行评价这一问题.事实上,是否具有评价所发现的规则的准确度的能力无论是对知识发现产品的开发者还是对一般的用户来说都是很重要的,这是因为:对于开发者来说,可以将自己的产品与同类产品进行比较,从而进行不断的改进和提高;对一般用户来说,猜测误差低的方法所挖掘出的规则抓住了一个数据集的更多特征,从而可以更可信地猜测丢失的值.

2.2 基于距离的关联规则

2.2.1 定 义

现在我们给出基于距离的关联规则的定义:

定义 2.1. 设 $X_1,X_2,\ldots,X_x,Y_1,Y_2,\ldots,Y_y$ 是两两互不相交的属性集,我们称式(2.2)是一条基于距离的关联规则:

$$C_{X1}C_{X2}\ldots C_{Xx}\Rightarrow C_{Y1}C_{Y2}\ldots C_{Yy}. \tag{2.2}$$

其中 C_{Xi} 和 C_{Yi} 分别是数据集在属性集 X_i 和 Y_i 上的聚类,如果

$$D(C_{Yj}[Y_j],C_{Xi}[Y])\leqslant D_0,\quad 1\leqslant i\leqslant x,1\leqslant j\leqslant y, \tag{2.3}$$

$$D(C_{Xi}[X_i],C_{Xj}[X_i])\leqslant d_0^{Xi},\quad \forall\, i\neq j, \tag{2.4}$$

$$D(C_{Y_i}[Y_i],C_{Y_j}[Y_i])\leqslant d_0^{Y_i},\quad \forall\, i\neq j, \tag{2.5}$$

其中 D_0 代表关联度,$d_0^{X_i}$代表规则前件中聚类之间的接近度阈值,$d_0^{Y_i}$ 代表规则后件中聚类之间的接近度阈值.

基于距离的关联规则要求规则前件中每个聚类之间密切相关,规则后件中每个聚类之间密切相关,并且前件中的每个聚类和后件中的每个聚类之间两两强关联.

在以上定义中,规则前件和后件都可以包含任意数目的聚类.我们称之为 $N:N$ 基于距离的关联规则.去除式(2.3),即只有一个后件,便可得到 $N:1$ 基于距离的关联规则.去除式(2.4)和式(2.5),即只保留一个前件和一个后件,便可得到 1:1 基于距离的关联规则.在应用中,我们可以根据不同的需要分别或混合选用这 3 种规则形式.

在以上定义中,可以认为关联度 D_0 对应于经典关联规则的可信度,接近度 $d_0^{X_i}$和 $d_0^{Y_i}$对应于经典关联规则的支持度.

2.2.2 基于距离的关联规则的挖掘过程

(1) 聚类数据

这里,我们采用 BIRCH(balanced iterative reducing and clustering using hierarchies)聚类工具[9]对数据集进行聚类,采用的距离为曼哈顿距离 $D(C1[X],C2[X])=|\overrightarrow{X0_1}-\overrightarrow{X0_2}|$,其中 $X0_1$ 和 $X0_2$ 分别为聚类 $C1[X]$和聚类 $C2[X]$的重心.BIRCH 的一个重要特性就是对大数据集来说,它能在一定的时空资源约束下,随着数据不断地被调入内存,渐增地对数据进行聚类.它只需对原始数据集进行一次遍历,并且采用了聚类特征(clustering feature,简称 CF)及聚类特征树这两个数据结构来描述所形成的聚类.一个聚类的聚类特征被定义为一个三元组,即 CF=$(N,\overrightarrow{LS},SS)$,其中 N 是此聚类中的数据点数;$\overrightarrow{LS}$ 是这 N 个数据点的线性和,即 $\sum_{i=1}^{N}\vec{X}_i$;SS 是这 N 个数据点的平方和,即 $\sum_{i=1}^{N}\vec{X}_i^2$.一个聚类特征树是一棵高度平衡的树,它含有两种类型的结点:叶结点和非叶结点以及两种参数:分枝因子(对非叶结点为 B,对叶结点为 L)和阈值 T.一个非叶结点至多含有 B 个$[CF_i,\text{childs}_i]$项,其中 $i=1,2,3,\ldots,B$,child_i 是一个指向它的第 i 个孩子的指针.所以,一个非叶结点描述了一个聚类,这个聚类由它含有的全部项所描述的子聚类组成.一个叶结点至多含有 L 个$[CF_i]$项,其中 $i=1,2,3,\ldots,L$,一个 CF_i 是它的第 i 个子聚类的 CF.所以一个叶结点也描述了一个聚类,这个聚类由它含有的全部项所描述的子聚类组成.此外,每个叶结点还含有两个指针,分别指向它的前驱和后继,一个叶结点中的所有项都必须满足阈值要求 T,这里,我们将此阈值要求定义为:叶结点中每一项的直径都不大于 T.

在基于距离的关联规则的挖掘算法中,对聚类特征进行了扩充,即除了保留前述的聚类特征以外,还保留了聚类在其他属性上的投影,即对所有的属性集 $Y\neq X$:

$$\sum_{i=1}^{N}t_i[Y],\ \sum_{i=1}^{N}t_i[Y]^2. \tag{2.6}$$

我们称此扩充表示的聚类特征为关联聚类特征(association clustering feature,简称 ACF).相应地,一个关联聚类特征树就是聚类特征树的扩充,它的叶结点描述了关联聚类特征,它的非叶结点则仍然描述聚类特征.

(2) 从数据的聚类中发现关联规则

在聚类阶段选出了支持度大于给定阈值的频繁聚类,现在我们来讨论如何从这些频繁聚类中挖掘出关联规则.为此,我们先给出聚类图的概念.

定义 2.2. 一个聚类图的一个结点 n_c 对应于数据集的一个聚类 c;从结点 n_{cx} 到 n_{cy} 存在一条边,如果 $D(C_X[X],C_Y[X])\leqslant d_0^X$ 并且 $D(C_X[Y],C_Y[Y])\leqslant d_0^Y$.

这样,给定一个数据集所有聚类的聚类关联特征,我们就可以利用某一距离度量标准,例如欧几里得或曼哈坦距离计算出这些聚类的聚类图.然后再从这些聚类图中,找出所有最大完全子图(maximal cliques,即图中的任意两个顶点之间都有一条边).一个最大完全子图和它的任意子集,都可以构成一条基于距离的关联规则的前件或后件.

设 C_1 和 C_2 分别是相应于两个聚类的集合 $C_X=\{C_{X1},C_{X2},\ldots,C_{Xx}\}$,$C_Y=\{C_{Y1},C_{Y2},\ldots,C_{Yy}\}$ 的最大完全子图,$\text{assoc}(C_{Yj})=\{C_{xi}|D(C_{Yj}[Y_{j-}],C_{Xi}[Y_j])\leqslant D_0,1\leqslant j\leqslant y,1\leqslant i\leqslant x\}$,对所有两两最大完全子图,重复以下过程:对所有的 $s\in\text{assoc}(C_{Yj})$,$s\Rightarrow C_{yj}$ 是一条关联规则;如果 C_X 的一个聚类的集合 $s'\in\text{assoc}(C_{Yj})$ 且 $s'\in\text{assoc}(C_{Yk})$,则 $s'\rightarrow C_{yj}C_{Yk}$ 也是一条关联规则.设 $C_Y'\subseteq C\supset Y$ 是一个最大完全子图 C_2 的子集,则对于最大完全子图 C_1 的每一个子集 $C_X'\subset C_X$,如果 $C_X'\subseteq\bigcap_{C_{yj}\in C_Y'}\text{assoc}(C_{yj})$,则存在规则 $C_X'\Rightarrow C_Y'$.

2.3 发现算法

在以上讨论的基础上,现在我们提出一个不完全数据库中的知识发现算法.

不妨考虑一个给定的含有丢失数据的事务数据矩阵.在从磁盘上读入数据的过程中,每当遇到带有丢失值的事务时,先将此事务放入一个磁盘存储区,这样,在分出了带有丢失值的事务之后,再基于不带有丢失值的事务,计算出相应的协方差矩阵,然后求出特征值和特征向量,进而得出线性关联规则以猜测出丢失的值.

在猜测出丢失的数据得到相应于原先数据矩阵的完全数据矩阵以后,就进入基于距离的关联规则的挖掘过程.首先是利用以上所介绍过的 BIRCH 聚类工具对完全数据矩阵中的数据进行聚类,聚类的结果用关联聚类特征树来描述.随着聚类的进行,关联聚类特征树被逐步地建立.在完成了对数据矩阵中数据的聚类之后,再计算出相应的聚类图,然后从聚类图中找出最大完全子图.最大完全子图中的节点就是基于距离的关联规则中的频繁项集.因此,一个最大完全子图和它的任意子集都可以作为一条基于距离的关联规则的前件或后件.

2.3.1 算法描述

(1) 读入数据矩阵并分离出带有丢失数据的事务(由此得到一个由已知数据构成的完全矩阵).

(2) 通过线性关联规则猜测出丢失的数据.

(2.1) 计算由第(1)步获得的完全矩阵的协方差矩阵;

(2.2) 计算协方差矩阵的特征值和特征向量;

(2.3) 选择 K 个特征向量,确定线性关联规则;

(2.4) 利用线性关联规则猜测丢失的值(由此得到一个相应于原先矩阵的完全矩阵).

(3) 运用基于距离的方法挖掘关联规则.

(3.1) 对所得完全矩阵中的数据进行聚类;

(3.2) 从数据的聚类中挖掘关联规则;

(3.2.1) 计算聚类图,

(3.2.2) 计算聚类图的最大完全子图,

(3.2.3) 基于最大完全子图挖掘关联规则.

2.3.2 算法的复杂度

设 N,M 分别为事务数据矩阵的行和列.步骤(1)需要扫描数据矩阵1次,复杂度为 $O(NM)$.步骤(2.1)中计算协方差矩阵的复杂度为 $O(NM^2)$.计算特征值和特征向量的复杂度为 $O(M^3)$.第(2.3)步含有一个排序.第(2.4)步主要涉及矩阵的乘法运算.因此,当 $N \gg M$ 时,步骤(2)的复杂度为 $O(NM^2)$.对于步骤(2.3.1),利用 Birch 聚类工具进行数据的聚类,这时需要重新扫描数据矩阵1次,矩阵的每一行数据需要沿着一棵关联聚类特征树遍历至多 $\mathrm{Log}_L N$ 步,其中 L 为聚类特征树内部结点的分枝因子.这样,对于整个数据矩阵,聚类数据的时间复杂度为 $O(MN(\mathrm{Log}_L N))$.值得一提的是,有时所建立的关联聚类特征树太大以致于不能驻留在内存,这时就必须调整树的相关阈值,以便对所形成的聚类特征树进行重建,重建的复杂度为 $O(M(\mathrm{Log}_L N)^2)$.所以,在最坏的情况下,聚类数据的时间复杂度为 $O(M(\mathrm{Log}_L N)_2)$.在第(3.2)步中,必须考虑所有可能的两两聚类组合以构造聚类图,并且还要找出全部的最大完全子图,再挖掘出关联规则,如果不采用启发式知识,则复杂度与聚类的个数呈指数增长关系.因此,在实际应用中,必须根据问题领域确定出合适的启发式知识来指导该步的进行.

3 实验结果

我们分别在 Solaris 的 UNIX 环境下和 PC 机上用 C 语言实现了以上的算法.在进行猜测丢失数据的实验时,我们是先用一组不含有丢失数据的训练数据求出所需的主成分,然后再对测试数据中的丢失值进行猜测.对于如表1所示的测试数据(取自文献[10],第425页)中的丢失值(以黑体字表示),我们得到了如表2所示的猜测结果.表3是利用列平均值法(即用一列的平均值替代出现在该列中的丢失值)来猜测丢失值的结果.

Table 1 Test data

表1 测试数据

Rec #	$X1$	$X2$	$X3$	$X4$
1	**149.3**	4.2	108.1	15.9
5	180.8	**1.1**	132.1	18.8
7	202.1	2.1	146.0	**22.7**
9	226.1	5.0	**162.3**	**28.1**
11	239.0	0.7	**167.6**	26.3

Table 2 Guessed missing data by principal component analysis

表2 以主成分法测试丢失值的结果

Rec #	$X1$	$X2$	$X3$	$X4$
1	**155.18**	4.2	108.1	15.9
5	180.8	**0.33**	132.1	18.8
7	202.1	2.1	146.0	**30.16**
9	226.1	5.0	**155.38**	**33.19**
11	239.0	0.7	**163.38**	26.3

Table 3 Guessed missing data by column average

表3 以列平均值猜测丢失的值

Rec #	$X1$	$X2$	$X3$	$X4$
1	**194.59**	4.2	108.1	15.9
5	180.8	**3.3**	132.1	18.8
7	202.1	2.1	146.0	**21.89**
9	226.1	5.0	**139.74**	**21.89**
11	239.0	0.7	**139.74**	26.3

从表 2 和表 3 我们不难发现:主成分分析法的猜测准确度比列平均值法的猜测准确度要高.此外,由式(2.2)我们也可以得出:主成分法的猜测误差为 4.146,而列平均值法的猜测误差为 76.827.

在挖掘基于距离的关联规则的实验中,我们首先利用 Birch 对数据聚类,共形成了 25 个聚类,分别用 1,2,...,25 来代表.根据这 25 个聚类的关联聚类特征,并且我们取关联度阈值为 60,前件中聚类之间的接近度阈值为 10,后件中聚类之间的接近度阈值 50,由此计算出相应的聚类图见表 4,其中数字代表各个聚类.

Table 4 Computed cluster graph
表 4 计算出的聚类图

Cluster①	Associated cluster②	Cluster	Associated cluster	Cluster	Associated cluster	Cluster	Associated cluster	Cluster	Associated cluster
1	6,11	6		16,23	11	16,21,24	16	21	9,11,13,15
2	6,11,12	7	11,14, 21,25	12		17	25	22	18
3	9,19,20, 23	8	13,15,17, 18,24	13	16,21	18	22,24	23	3,6
4	15,20, 24	9	11,13,15, 21,25	14	16,17, 21,24	19	24	24	4,8,11,14,15
5	25	10		15	16,21	20	24	25	5,7,9,17

①聚类,②关联的聚类.

从此聚类图中,进一步找出了 35 个最大完全子图,见表 5.例如,最大完全子图 4 20 24 表示第 4、第 20、第 24 个聚类构成了一个三角形,即每两个顶点之间都有一条边.最后,从这 35 个最大完全子图中,我们发现了 90 多条规则.例如,我们发现最大完全子图 4 20 24 分别和最大完全子图 8 13 和最大完全子图 8 15 相关联,即有基于距离的关联规则:{4 20 24}⇒{8 13}和{4 20 24}⇒{8 15}成立.

Table 5 Computed maximal cliques
表 5 计算出的最大完全子图

No.①	Maximal cliques②	No.	Maximal cliques	No.	Maximal cliques	No.	Maximal cliques	No.	Maximal cliques
1	4 20 24	8	14 16	15	9 11 21	22	7 25	29	3 20
2	19 24	9	14 17	16	11 24	23	6 16	30	3 23
3	18 22	10	7 14 21	17	9 25	24	6 23	31	2 6
4	8 18 24	11	14 24	18	8 13	25	5 25	32	2 11
5	17 25	12	13 16	19	8 15	26	4 15	33	2 12
6	15 16	13	9 13 21	20	8 17	27	3 9	34	1 6
7	9 15 21	14	11 16	21	7 11 21	28	3 19	35	1 11

①序号,②最大完全子图.

从理论上说,从聚类图中求最大完全子图是一个 NP-完全问题.对此已有不少研究者提出了一些算法[11,12],这些算法能够进一步降低计算最大完全子图的时空资源.在我们的实验中,从聚类图中计算最大完全子图的时间约为 4 秒.并且,如果所求得的聚类图较为稀疏,则最大完全子图可以通过枚举求得.另外,还可以针对具体问题领域采用启发式的方法(例如,可以忽略数据点稀疏的聚类)来求最大完全子图.

4 与相关工作的比较

对于不完全数据库中关联规则的发现问题,目前的处理方法之一是直接去除带有丢失属性值

的元组[4].这一方法的缺陷就是在去除了带有丢失属性值的元组以后,我们可能得不到一个足够大的数据库,从而失去了原先数据库中的许多有用信息.在本文的方法中,我们不是去除带有丢失数据的元组,而是在保留它们的同时,直接猜测出丢失的属性值.另一方面,在经典的关联规则中,我们运用支持度和可信度这两个指标来度量所挖掘的关联规则.在挖掘经典的关联规则时,用户必须预先确定支持度阈值和可信度阈值.这样就出现了一个问题,即用户难以判断基于此预先确定的支持度和可信度阈值究竟是否能从所给数据库中挖掘出上百乃至上千条规则,还是只能得出几条规则甚至连一条规则也得不出呢?导致这一问题的原因是,因为经典的支持度和可信度所基于的是集合成员隶属关系,是对数据库元组间比例关系的度量,而不是对数据值的度量.因此,在本文的方法中,我们利用数据之间的距离作为挖掘关联规则的兴趣度量标准,这样,一方面较好地利用了原始数据中的数量信息,另一方面,也避开了经典的支持度和可信度所固有的不足.

5 结束语

数据的不完全已经严重影响了知识发现的理论和方法在现实世界应用的效果和进程,研究如何在数据不完全的环境下进行知识发现无疑是一个重要的研究课题.本文讨论了在部分数据丢失的情形下如何猜测丢失的数据值,然后在此基础上发现基于距离的关联规则的方法,并给出了实验结果,同时还与其他方法进行了比较.进一步的研究工作是考虑在同时存在连续形属性、离散型属性和分类型属性的不完全数据库中如何猜测丢失的数据和挖掘基于距离的关联规则,这将涉及应用于不同类型数据的兴趣度量标准的结合问题.

References:

[1] Agrawal, R., Mannila, H., Srikant, R., *et al.* Fast discovery of association rules. In: Fayyad, U., Piatetsky-Shapiro, G., Smyth P., *et al.*, eds. Advance in Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI Press, 1995. 307~328.

[2] Srikant, R., Vu, Q., Agrawal, R. Mining association rules with item constraints. In: Heckerman, D., Mannila, H., Pregibon, D., *et al.*, eds. Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI Press, 1997. 67~73.

[3] Agrawal, R., Shafer, J.C. Parallel mining of association rules. IEEE Transactions on Knowledge and Data Engineering, 1996,8(6):962~969.

[4] Ragel, A., Cremilleux, B. Treatment of missing values for association. In: Wu, Xin-dong, Ramamohanarao, K., Korb, K.B., eds. Research and Development in Knowledge Discovery and Data Mining. New York: Springer-Verlag, 1998. 258~269.

[5] Kononenko, I., Bratko, I. Experiment in automatic learning of medical diagnostic rules. Technical Report, Ljubljana, Yugoslavia: Jozef Stefan Institute, 1984.

[6] Quinlan, J.R. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann Publishers, Inc., 1993.

[7] Lakshminarayan, K., Harp, S., Goldman, R., *et al.* Imputation of missing data using machine learning techniques. In: Simoudis, E., Han, J., Fayyad, U., eds. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI Press, 1996. 140~145.

[8] Jolliffe, I.T. Principal Component Analysis. New York: Springer-Verlag, 1986.

[9] Zhang, T., Ramakrishnan, R., Livny, M. BIRCH: an efficient data clustering method for very large databases. Technical Report, Department of Computer Sciences, University of Wisconsin-Madison, 1995.

[10] Sun, Wen-shuang, Chen, Lian-xiang. Multi-Statistical Analysis. Beijing: Higher Education Publishing Company, 1994 (in Chinese).

[11] Chiba, N., Nishizeki, T. Arboricity and subgraph listing algorithms. Journal of Computing, 1985,14(1):210~223.

[12] Tsukiyama, S., Ide, M., Ariyoshi, H., *et al.* A new algorithm for generating all the maximal independent sets. Journal of Computing, 1977,6(3):505~517.

附中文参考文献:

[10] 孙文爽,陈兰祥.多元统计分析.北京:高等教育出版社,1994.

An Approach for Knowledge Discovery under the Environment of Incomplete Data*

WANG Qing-yi, CAI Zhi, ZOU Xiang, CAI Qing-sheng

(*Department of Computer Science, University of Science and Technology of China, Hefei* 230027, *China*)

E-mail: qywang@ustc.edu.cn

http://www.ustc.edu.cn

Abstract: The current researches of knowledge discovery are introduced under the environment of incomplete data, and then an approach is presented for knowledge discovery in incomplete databases through the two parts. Firstly, the method of how to guess the missing data is in detail discussed and the definition as well as the mining method of distance-based association rule is given. Then based on this, an algorithm is described in detail for discovering knowledge in incomplete databases, the complexity of the algorithm is analyzed and the experimental results are also given. The paper finally concludes with a comparison of the approach proposed in the paper with other related ones.

Key words: missing data; incomplete database; knowledge discovery; cluster; distance-based association rule

* Received January 13, 2000; accepted June 26, 2000

Supported by the National Natural Science Foundation of China under Grant No. 69[illegible]