发掘多值属性的关联规则*

张朝晖 陆玉昌 张 钹

(清华大学计算机科学与技术系 北京 100084)
(清华大学智能技术与系统国家重点实验室 北京 100084)

摘要 属性值可以取布尔量或多值量.从以布尔量描述的数据中发掘关联规则已经有比较成熟的系统和方法,而对于多值量则不然.将多值量的数据转化为布尔型的数据是一条方便、有效的途径.提出一种算法,根据数据本身的情况决定多值量的划分,进而将划分后的区段映射为布尔量.在此基础上可发掘容易理解且具有概括性的、有效的关联规则.

关键词 数据采掘,关联规则,聚类算法.

中图法分类号 TP311

当今世界,数据每天都在迅猛地增长.据估计,全世界的信息量每 20 个月翻一番.人们保存如此大量的数据,一是因为计算机技术的发展使之变得方便可行,二是因为这些数据有巨大的潜在作用.然而,如何有效地使用这些数据却成为一个问题,因为常常是数据丰富而知识缺乏,利用当前的数据库技术并不能很好地发挥这些数据的作用.

数据采掘(Data Mining)是数据库中知识发现 KDD(knowledge discovery in databases)的核心,它为大量数据的利用提供了有效的工具.自从 1989 年第 1 届 KDD 专题研讨会举办以来,数据采掘的研究方兴未艾.从 1995 年开始,每年举办一次的 KDD 国际会议,将 KDD 方面的研究推向了高潮.KDD 可以定义如下[1]:从数据中得出新的、有效的、有潜在用途的、可理解的模式的非平凡过程.

关联规则[2]是当前数据采掘研究的主要模式之一,侧重于确定数据中不同领域之间的联系,找出满足给定支持度和可信度阈值的多个域之间的依赖关系.下面是一个直观的关联规则的例子:在计算机配件商店中,70%的包含键盘的交易中包含鼠标,在所有交易中,有 6%同时包含这两种物品.规则表示为

键盘⇒鼠标 (可信度 70%,支持度 6%)

关联规则可以分为两种:布尔型关联规则和多值关联规则.[3]许多文献[2,5~8]都讨论了发掘布尔型关联规则问题[4],BARP(Boolean association rules problem),它可以看作是发掘多值关联规则问题 QARP(quantitative association rules problem)的基础和特例,是在属性值为布尔量的关系表中寻找属性值为"1"的属性之间的关系.多值属性可分为数量属性(Quantitative Attribute),如年龄、价格等;类别属性(Categorical Attribute),如品牌、制造商等.

QARP 比较复杂,一种自然的想法是将它转换为 BARP.当全部属性的取值数量都是有限的时候,只需将每个属性值映射为一个布尔型属性即可.当属性的取值范围很宽时,则需将其分为若干区段,然后将每个区段映射为一个布尔型属性.

于是,如何划分区段是实现 QARP 到 BARP 转变的关键.这里面有两个互相牵制的问题:当区段的范围太窄时,则可能使每个区段对应的属性的支持度很低,而出现"最小支持度问题";当区段的范围太宽时,则可能使每个区段对应的属性的可信度很低,而出现"最小可信度问题".

一种简单直观的方法是将属性值区域相等地划分成区段[3],但这种方法得出的划分不能很好地表示数据的分布,特别是当属性值分布不均匀的时候.本文提出一种聚类算法,根据数据库中数据的分布情况决定属性值如何划分区段,并可将相关的区段进行合并.在此基础上发掘得到的多值关联规则可具有有效性和可理解性.

---

* 本文研究得到国家自然科学基金和国防预研基金资助.
作者张朝晖,1970 年生,博士生,主要研究领域为数据采掘,机器学习,神经网络.陆玉昌,1937 年生,教授,主要研究领域为数据挖掘,知识发现,机器学习,知识工程.张钹,1935 年生,教授,博导,中国科学院院士,主要研究领域为人工智能,计算机应用.

本文通讯联系人:陆玉昌,北京 100084,清华大学计算机系智能技术与系统国家重点实验室

本文 1997-06-12 收到原稿,1997-10-23 收到修改稿

1 关联规则

从数据库中发掘的规则可以有以下几种:特征规则、区分规则、聚类规则、关联规则和进化规则等.关联规则是比较新的一种,由 R. Agrawal 于 1993 年提出.[2]

令 $I=\{i_1,i_2,i_3,\ldots,i_m\}$ 为项的集合,D 称为交易的集合,D 中每个交易 T 为项的集合,即 $T\subseteq I$.

定义 1. 如果对于 I 中一些项的集合 X 有 $X\subseteq T$,则称 T 包含 X.

定义 2. 一条关联规则是如下形式的蕴涵式 $X\Rightarrow Y$,这里,$X\subset I$,$Y\subset I$ 且 $X\cap Y=\varnothing$.规则 $X\Rightarrow Y$ 在交易集合 D 中成立,如果 D 中有 $s\%$的交易包含 $X\cup Y$,且 D 中有 $c\%$的包含 X 的交易也包含 Y.这里,s 称为支持度,c 称为可信度.

定义 3. 发掘关联规则问题就是在给定的交易集合 D 中产生所有满足最小支持度(MinSupp)和最小可信度(MinConf)的关联规则的过程.

发掘关联规则问题可以分为两个子问题.

(1) 寻找所有这样的项的集合(Itemsets),它们的支持度超过用户给定的最小支持度.这个项的集合称为频繁集(Frequent Itemset).

(2) 应用频繁集产生规则.一般的想法是,如果 $ABCD$ 和 AB 是频繁集,那么,可以通过计算可信度 $conf=supp(ABCD)/supp(AB)$ 来确定规则 $AB\Rightarrow CD$ 是否成立.当可信度 $conf\geq$最小可信度时,规则成立.其中 $supp(X)$ 表示 X 的支持度.

随着关联规则越来越受到重视,许多算法和系统被相继提出[3~7],大多集中于处理第 1 个子问题,其中产生频繁集的快速 Apriori 算法[5]被广泛采用.

多值关联规则问题和基本的关联规则问题略有差别,这里项的集合变为 $I_r=I\times P\times P$,即 $I_r=\{(x,l,u)|x\in I,l\in P,u\in P,l\in u,l\leq x\leq u\}$.$(x,l,u)\in I_r$ 表示属性 x 取值在 l 和 u 之间.其中 I 为属性集合,P 为正整数集合.对于任何 $X\subseteq I_r$,$attribute(X)=\{x|(x,l,u)\in X\}$.

定义 4. 如果 $\forall(x,l,u)\in X$,$\exists(x,v)\in T$,有 $l\leq v\leq u$,那么,称交易 $T(T\in D)$ 支持 $X(X\subseteq I_r)$.

定义 5. 多值关联规则是具有 $X\Rightarrow Y$ 这样的蕴涵式,其中 $X\subset I_r$,$Y\subset I_r$,且 $attribute(X)\cap attribute(Y)=\varnothing$.如果 D 中有 $s\%$的交易支持 $X\cup Y$,且 $c\%$的支持 X 的交易也支持 Y,则该规则的支持度和可信度分别为 s 和 c.

定义 6. 发掘多值关联规则问题就是在给定的交易集合 D 中产生所有满足最小支持度和最小可信度的多值关联规则的过程.

2 算法描述

我们采用将 QARP 映射为 BARP 的方法发掘多值关联规则.下面先给出发掘多值关联规则的总的算法 MAQA (mining associations among quantitative attributes),然后详细讨论其中的聚类算法和合并算法.

2.1 MAQA 算法

MAQA 算法将多值关联规则问题转化为布尔型关联规则问题,然后利用已有的发掘布尔型关联规则的方法得到有价值的规则.若属性为类别属性,则先将属性值映射为连续的整数,并使意义相近的取值相邻编号.

算法简述如下.

Step 1 对于多值属性 A,取值范围为 $[l,r]$.若为数量属性,则应用聚类算法 CP(clustering for partitioning)确定多值属性的划分;若为类别属性,则进行归纳划分.

Step 2 将划分后的属性区段 $[l_k,r_k]$ 或属性值映射成序对 $\langle A,k\rangle$,进而映射为布尔属性 A_m,所有这样的属性构成项集.

Step 3 从项集中寻找所有有价值的项,构成频繁项集;有价值的项是指支持它的交易的数量超过给定的 Min*Supp* 的项.

Step 4 在频繁集中迭代地搜索出组合后的支持度超过给定阈值的两个项,并将其组合并加入频繁集中;如果是相同属性的相邻区段,则进一步合并.

Step 5 应用频繁集产生关联规则.如果 $ABCD$ 和 AB 都是频繁集,则判定规则 $AB\Rightarrow CD$ 是否成立,是通过计算可信度 $conf=supp(ABCD)/supp(AB)$ 是否超过最小可信度来决定.如果超过,则规则成立.其中 $supp(X)$ 表示 X 的支持度.

Step 6 确定有价值的(Interesting)关联规则作为输出.

上面的算法中,Step 2 和 Step 5 非常直观简单,Step 3 采用前面提到的 Apriori 算法,Step 6 采用典型文献[5]中的interest度量方法,下面详细描述 Step 1 和 Step 4.

2.2 确定多值属性划分的聚类算法 CP

如果一个数量属性的取值数目过多时,则将这个属性划分为若干个区段.如果一个类别属性的取值过多时,则将

其属性值进行归纳，例如，将铅笔、橡皮和钢笔归纳为文具，然后将能归纳在一起的属性值划分为同一个区段．下面主要讨论数量属性的情况，其中属性均指数量属性．

对于一个属性的每一个取值，计算 D 中含有该属性值的交易数，结果构成集合 I_x．$I_x=\{(x,v,n)\mid x\in I,v$ 是属性 x 的一个取值，$n\in P$ 是 D 中交易的个数$\}$．对于任何属性 x，我们将其对应的集合 I_x 映射为如图 1 所示的二维图．x 轴表示属性的取值，y 轴表示 D 中所对应的交易数．如果将图中的点用曲线连接起来，则得到图 2．因为各个取值的交易数不同，所以曲线是波动的．

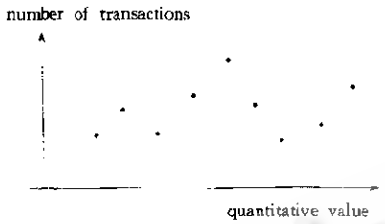

图1　交易数分布的点图

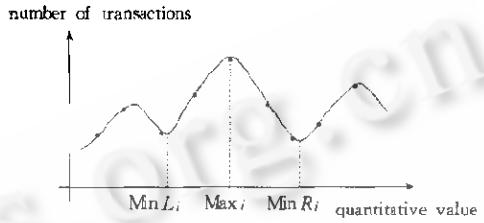

图2　交易分布的曲线图

对应于每个局部最大点 Max$_i$，存在两个局部最小点 MinL_i 和 MinR_i．将 MinL_i 和 MinR_i 之间的交易数相加得到 Sum_i，再将所有的 K 个 Sum_i 相加得到 S．将 Min$Supp$ 设为 $c\cdot S_{ave}$，这里 $S_{ave}=S/K$，$c>0$ 为与问题有关的常数．当 Sum_i 大于 $c\cdot S_{ave}$ 时，则认为 MinL_i 和 MinR_i 之间的区段是有价值的．这里有 3 个问题：

(1) S_{ave} 的代表性．如果某个 Sum_i 特别大和/或某个 Sum_j 特别小，则会影响平均值的代表性．当 Sum_i 的值随机分布时，S_{ave} 不应该接近 $\max\{Sum_i\}$ 或 $\min\{Sum_i\}$．

(2) 信息丢失．如果我们将交易数小于 S_{ave} 的区段全部丢弃，则会有许多信息丢失．虽然交易数可能很少，但可能能与相邻的区段合并构成一个大的有价值的区段．

(3) 可区分性．如果每个区段的交易数都差不多，使每个 Sum_i 都接近 S_{ave}，于是很难区分哪一个区段更有价值．

对于第 1 个问题，可以用如下方法解决：在计算局部最大之前，从 S 中减去 $\max\{Sum_i\}$ 和 $\min\{Sum_i\}$ 两个值，并从 K 中减去 2，这样使平均值更具代表性．

对于第 2 个问题，我们采用合并相邻区段的方法．如果一个区段的宽度很窄，与相邻区段距离很近，而且它们之间的最低点与它们之中较低的局部最大点的差值很小，那么，将这两个区段合并．图 3 给出了这种方法的直观含义．

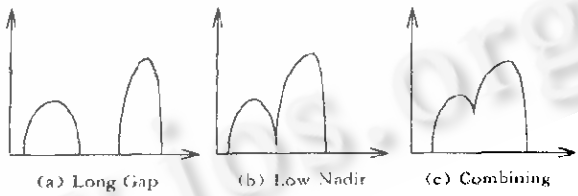

图 3　合并相邻区段的条件

对于第 3 个问题，我们采用合并相邻区段，并同时考虑宽度的方法来解决．宽度通过计算交易数和属性值的比率体现出来．如果 $Sum_i/(\text{Min}R_i-\text{Min}L_i)>S/(\text{Min}R-\text{Min}L)$，那么，我们认为 Sum_i 所对应的区段是有价值的．这里，Min$R$ 和 Min$L$ 分别是该属性所能取的最大值和最小值．

聚类算法可简述如下．

for 每个取值数$>N$ 的属性 do

Step 1 计算每个属性值所对应的交易数 C_i；

Step 2 寻找所有的局部最大点 Max$_i$ 和最小点 MinL_i 和 MinR_i 来确定区段；

Step 3 计算每个 MinL_i 和 MinR_i 之间的交易数 Sum_i；

Step 4 如果满足合并条件则合并两个相邻区段，得到 K 个区段；

Step 5 $S=\sum Sum_i-\max\{Sum_i\}-\min\{Sum_i\}$；

Step 6 $S_{ave}=S/(K-2)$；

Step 7 寻找所有大于 $c \cdot S_{ave}$ 的 Sum_i，并将结果存于 S_{tmp}；

Step 8 For S_{tmp} 中每个区段 j do

if $Sum_j/(\mathrm{Min}R_j-\mathrm{Min}L_j)>S/(\mathrm{Min}R-\mathrm{Min}L)$

then 保存区段 j 于 S_{res}

结果为有价值的区段，保存在 S_{res} 中.

2.3 合并数量属性的相邻值

第 2.2 节讨论了为达到一定的支持度而进行的同一属性的合并，我们还对一些达到了最小支持度 Min*Supp* 的区段进行合并，即合并频繁集中同一属性的相邻区段.

为了说明合并相邻区段的原因，考虑下面的规则：

$\langle Age=20..25\rangle\Rightarrow\langle Cars=1..2\rangle$ （支持度 3%，可信度 70%）

$\langle Age=26..30\rangle\Rightarrow\langle Cars=1..2\rangle$ （支持度 4%，可信度 70%）

在这种情况下，可以将两条规则合并为 1 条：

$\langle Age=20..30\rangle\Rightarrow\langle Cars=1..2\rangle$ （支持度 7%，可信度 70%）

可以看出，这条规则更有概括性. 虽然在这个转换过程中丢失了一些信息，即小区段的支持度不再能表达出来，但是，这却使规则数减少，使规则集简化了. 即使两条规则的可信度稍有不同，信息损失也不大. 这就是合并频繁集中的 $(Age,20,25)$ 与 $(Age,26,30)$ 两个元素为 $(Age,20,30)$ 的原因.

还有一个问题就是合并后是否删除原来的两项. 仍然用上面的例子，因为 $C=(Age,20,30)$ 覆盖了 $A=(Age,20,25)$ 与 $B=(Age,26,30)$ 两个元素，即 A 或 B 成立时，C 也一定成立. 而且多值关联规则 $X\Rightarrow Y$ 成立要满足 $attribute(X)\cap attribute(Y)=\varnothing$，所以，不会出现 $B\Rightarrow A$ 或 $A\Rightarrow B$ 这样的规则. 这与布尔关联规则问题不同，当 PQ，P 和 Q 同时存在于频繁集中时，可能会得到 $P\Rightarrow Q$ 或 $Q\Rightarrow P$ 这样的规则. 于是，似乎可以删除 A 或 B. 但是，当考虑规则的可信度时则不能删除. 例如，假设有如下的规则：

(1) $\langle Age=20..25\rangle\Rightarrow\langle Cars=1..2\rangle$ （支持度 5%，可信度 60%）

(2) $\langle Age=20..25\rangle\Rightarrow\langle Cars=3..4\rangle$ （支持度 3%，可信度 36%）

(3) $\langle Age=26..30\rangle\Rightarrow\langle Cars=1..2\rangle$ （支持度 3%，可信度 36%）

(4) $\langle Age=26..30\rangle\Rightarrow\langle Cars=3..4\rangle$ （支持度 5%，可信度 60%）

合并后则有

(5) $\langle Age=20..30\rangle\Rightarrow\langle Cars=1..2\rangle$ （支持度 8%，可信度 48%）

(6) $\langle Age=20..30\rangle\Rightarrow\langle Cars=3..4\rangle$ （支持度 8%，可信度 48%）

若 MinSupp=3%，MinConf=50%，则若不删除频繁集中的两项 $(Age,20,25)$ 与 $(Age,26,30)$，可得到规则(1)和(4)，而删除后则得不到上面的任何一条规则.

因此，在 MAQA 算法的 Step 4 中，合并频繁集中的两个元素之后并不删除原来的元素，在迭代中则将合并的和未合并的元素同等处理. 生成规则后，如果有冗余，再删除冗余规则. 如上例中，在 MAQA 的 Step 4 的第 2 次迭代中不仅产生{{$(Age,20,25)$,$(Cars,1,2)$},{$(Age,26,30)$,$(Cars,1,2)$},{$(Age,20,25)$,$(Cars,3,4)$},{$(Age,26,30)$,$(Cars,3,4)$},{$(Age,20,30)$}}，还要产生元素{$(Age,20,30)$,$(Cars,1,2)$}和元素{$(Age,20,30)$,$(Cars,3,4)$}. 另外，若 $(Age,20,25)$ 和 $(Age,26,30)$ 作为 Apriori 算法中的 $p.item_{k-1}$ 和 $q.item_{k-1}$ 时，不进行合并.

3 总 结

本文分析了当前 KDD 的蓬勃发展形势，指出关联规则是 KDD 所要发掘的重要目标之一. 关联规则可分为布尔型关联规则和多值关联规则. 关于布尔型关联规则已经有许多方法和系统，而且多值关联规则也常常通过转化为布尔型关联规则来解决. 本文针对平均划分多值属性区段的缺陷，采用对类别属性进行概括、对数量属性进行聚类的办法将多值属性划分为区段，然后转化为布尔型属性，再利用发掘布尔型关联规则的方法发掘多值关联规则. 这种方式根据数据库中交易的分布决定属性值区段的划分，比平均划分的区段更有代表性，从而能得到更有效的规则. 文中给出了发掘多值关联规则的算法 MAQA 及聚类划分的详细步骤，并对多值关联规则特有的问题进行了分析，并给出了解决办法. 目前的工作还主要是方法上的研究，今后的工作则是从实际的气象数据库中发掘多值关联规则.

参考文献

1 Fayyad U, Shapiro G, Smyth P. The KDD process for extracting useful knowledge from volumes of data. Communications of

the ACM, 1996,39(11):27～34
2 Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD Conference on Management of Data. Washington D. C, 1993. 207～216
3 Srikant R, Agrawal R. Mining quantitative association rules in large relational tables. In: Proceedings of the ACM SIGMOD Conference on Management of Data. 1996
4 Zhang Zhao-hui, Lu Yu-chang, Zhang Bo. An effective partitioning-combining algorithm for discovering quantitative association rules. In: Proceedings of PAKDD. Singapore, World Scientific Publishing Co., 1997. 241～251
5 Agrawal R, Srikant R. Fast algorithm for mining association rules. In: Proceedings of the 20th International Conference on Very Large Databases. Santiago, Chile, 1994
6 Houtsma M, Swami A. Set-oriented mining of association rules. In: Proceedings of the 11th International Conference on Data Engineering. 1995. 25～33
7 Han J, Huang Y, Cercone C *et al*. Intelligent query answering by knowledge discovery techniques. IEEE Transactions on Knowledge and Data Engineering, 1996,8(3):373～390
8 Cheung D, Ng V, Fu A *et al*. Efficient mining of association rules in distributed databases. IEEE Transactions on Knowledge and Data Engineering, 1996,8(6):911～922

An Algorithm for Mining Quantitative Association Rules

ZHANG Zhao-hui　LU Yu-chang　ZHANG Bo

(*Department of Computer Science and Technology　Tsinghua University　Beijing*　100084)
(*State Key Laboratory of Intelligent Technology and Systems　Tsinghua University　Beijing*　100084)

Abstract　An attribute can be Boolean or quantitative. There are lots of systems and methods for mining Boolean association rules but few for quantitative. Mapping quantitative attributes into Boolean attributes is a convenient and efficient way. In this paper, a new clustering algorithm is presented. Quantitative attribute values are partitioned into intervals according to the distribution of them in database. Then the intervals are mapped into Boolean attributes. In this way, quantitative rules can be mined by using the techniques of mining Boolean association rules.

Key words　Data mining, association rules, clustering algorithm.