一种自适应的模糊关联规则挖掘算法

赵纪涛,马 莉,王现君,尚光龙
(河南大学 数据与知识工程研究所,河南 开封 475001)

摘 要:关联规则是数据挖掘的重要研究内容之一。传统的关联规则挖掘算法仅适于处理二元属性与分类属性。为更好地处理数量属性,提出了一种自适应的基于模糊概念的量化关联规则挖掘算法。该算法克服了传统的离散分区法的不足,改进了已有模糊关联规则支持度的计算方法。引入了一种基于聚类的隶属函数自动生成方法,使得模糊关联规则的发现不依赖于人类专家给出的隶属函数,使得关联规则的表示自然、简明,有利于专家理解。实验表明该算法是有效的。

关键词:模糊关联规则;模糊集;数据挖掘;自适应
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2008)05-0064-03

An Adaptive Algorithm for Mining Fuzzy Association Rules

ZHAO Ji-tao, MA Li, WANG Xian-jun, SHANG Guang-long
(Institute of Data and Knowledge Engineering, Henan University, Kaifeng 475001, China)

Abstract: Mining association rules is one of the important research problem in data mining, many algorithms have been proposed to find association rules in database with binary attribute and categorical attribute. Introduce an adaptive algorithm for mining fuzzy association rules. It overcomes the drawbacks caused by the traditional discrete interval method. The algorithm adopts an improved calculating measure of itemset. A method for automatic definition for membership function is proposed, which using fuzzy clustering from training example. The experimental results show that the algorithm is effective and can provide important mining results to users.

Key words: fuzzy association rules; fuzzy set; data mining; adaptive

0 引 言

Agrawal于1993年首次提出布尔型关联规则[1]问题,并提出Apriori算法。此后数据挖掘的研究者做了大量的工作,有的致力于算法的改进,提高算法的速度和有效性;有的提出新的算法[2]。这些算法对挖掘布尔型关联规则是非常有效的。

近年来,为处理现实中更常见的数值型数据,研究者的兴趣开始转向量化关联规则挖掘。采用划分区间,再组合近邻分区的方法,将量化关联规则转化为布尔关联规则挖掘。然而,该方法有显而易见的缺点,即容易忽略或过分强调分区临界值附近点在分区中所起的作用,这种划分也是一种硬划分。因此,模糊理论被引入到量化关联规则的挖掘中,提出了模糊关联规则挖掘算法[3~5]。该算法很好地解决了尖锐边界问题,但其采用"min(Set)"方法来求模糊频繁项目集支持度。因"min(Set)"是取小运算,结果只保留它们的下端信息,其余的信息全被舍弃。文中把min(Set)改为代数积,因为代数积在计算事务对项目集支持度时避免了取小运算的不足,同时,引入模糊隶属函数自动生成的算法。

收稿日期:2007-08-20
基金项目:国家自然科学基金资助项目(60474022);河南省高校杰出科研人才创新工程项目(2007KYCX018)
作者简介:赵纪涛(1982-),男,河南商丘人,硕士研究生,研究方向为数据挖掘、信息管理。

1 基本概念

对数据库的量化属性处理不采用区间划分法,而采用模糊概念对其进行抽象、概括,从而使得最终挖掘出的规则表示自然、易于专家理解。模糊概念的属性表示就是模糊集合论,其不明确的内涵和外延用隶属函数定量描述。设 $I=\{i_1,i_2,\cdots,i_n\}$ 是全体项目的集合,$T=\{t_1,t_2,\cdots,t_p\}$ 为全体事务的集合,每个事务 t_i 都有唯一的TID标识,且 $t_i\subseteq I$,$t_i[i_j]$ 表示事务 t_i 在属性 i_j 上的取值,属性 i_j 的域记为 $\mathrm{dom}(i_j)$。对 I 的任一数值属性 i_k,都有一个与之相对应的模糊集的集合 $F_{i_k}=\{f_{i_k}^1,f_{i_k}^2,\cdots,f_{i_k}^l\}$,模糊 f_{i_k} 集的隶属函数记为 $\mu(f_{i_k})$,其中 $k=1,2,\cdots,l$。

定义1[3]:模糊模式定义形式为 $P=A_1\wedge A_2\wedge\cdots\wedge A_k$,$A_i=(a_j,v_j)$,$a_j\in\mathrm{Attr}$,$v_j\in F_j(i=1,2,$

…,k),称模式 P 的长度为 k,A_i 称为项目。由 $A_1, A_2, \cdots, A_k$ 中,任意 $m(m < k)$ 个项目组成的模式称为 P 的子模式。对于事务 $T = \{t_1, t_2, \cdots, t_p\}$,$A_i = (a_j, f_{j_i})$,称 $\min(\{\mu(f_{j_i}(t_i)) \mid A_i = (a_j, f_{j_i}), (i = 1,2,\cdots, k)\})$ 为事务 T 对模糊模式 P 的支持,记为 $s(P,T)$。其中 min(Set) 表示集合 Set 中的最小元素。

上述定义在计算事务对项目集的支持度时采用 min(Set),这种算法的优点是运算简单,除不满足互补律外,与经典的集合运算十分相似。但也有缺点,因"∧"是取小运算,结果只保留它们的下端信息,其余的信息全被舍弃。文中把 min(Set) 改为代数积,因为代数积在计算事务对项目集支持度时避免了取小运算的不足。

定义 2:模式 $P = \{A_1 \wedge A_2 \wedge \cdots \wedge A_k\}$ 在 D 中的支持度 $\mathrm{Sup}(P) = \dfrac{\sum_{j}^{n} \prod_{m=1}^{k} t_j(f_m)}{n}$,规则 $A \Rightarrow B$ 的可信度 $\mathrm{Conf}(A \Rightarrow B) = \dfrac{\mathrm{Sup}(A \wedge B)}{\mathrm{Sup}(B)}$,即 A 发生时 B 发生的条件概率 $P(B \mid A)$,其中 A,B 均为模式。

为了挖掘关联有效的关联规则,必须定义最小支持度 minsup 和最小置信度 minconf。其中大于最小支持度的模式称为频繁模糊模式。挖掘关联规则即找出满足 $\mathrm{Sup}(A \wedge B) \geqslant \mathrm{minsup}$ 和 $\mathrm{Conf}(A \Rightarrow B) \geqslant \mathrm{minconf}$ 的规则 $A \Rightarrow B$。

定理 1:模糊模式 P 的支持度为 $\mathrm{Sup}(P)$,若 $\forall Q \subseteq P$,则有 $\mathrm{Sup}(Q) \geqslant \mathrm{Sup}(P)$;若 $\forall Q \supseteq P$,则有 $\mathrm{Sup}(R) \leqslant \mathrm{Sup}(P)$。

证明:由模糊模式的支持度定义,当函数取"min"或相乘时显然成立。

性质 1:如果模糊模式 P 是频繁的,$\forall Q \subseteq P$,则 Q 也是频繁的。

性质 2:如果模糊模式 P 是非频繁的,$\forall R \supseteq P$,则 R 也是非频繁的。

性质 1 和性质 2 可以用于超集的剪枝,这种基于支持度度量修剪指数空间的策略称为基于支持度的剪枝。这种剪枝策略依赖与支持度度量的一个关键性质,即一个项集的支持度绝不会超过它的子集的支持度。这个性质称为反单调性。

定义 3[6]:令 I 是项的集合,$J = 2^I$ 是 I 的幂集。度量 f 是单调的(或向上封闭的),如果

$\forall X, Y \in J: (X \subseteq Y) \rightarrow f(X) \leqslant f(Y)$;

另一方面,度量 f 是反单调的(或向下封闭的),如果

$\forall X, Y \in J: (X \subseteq Y) \rightarrow f(Y) \leqslant f(X)$。

2　挖掘算法与隶属函数的确定

因此可以对 Apriori 算法加以改造。整个挖掘过程分为两步:第一步挖掘频繁模糊模式,第二步由频繁模糊模式生成关联规则。与传统关联规则挖掘算法类似,其中第一步是挖掘的关键。下面就是生成频繁模糊模式的算法 FFP(Frequent Fuzzy Pattern):

输入:交易数据集 D;最小支持度 minsup;

输出:频繁模糊项目集 L;

$\mu_{R_j^{(k)}}: v_{ij} \rightarrow f^{(k)}, f_{ij}^{(k)} = \mu_{R_j^{(k)}}(v_{ij}), R_j^{(k)} = \sum_{i=1}^{n} \dfrac{f_{ij}^{(k)}}{v_{ij}}$ //数据变换

$\mathrm{count}_j^{(k)} = \sum_{i=1}^{n} f_{ij}^{(k)}$

$L_1 = \{R_j^{(k)} \mid \mathrm{count}_j^{(k)} \geqslant \mathrm{minsup}, 1 \leqslant j \leqslant n, 1 \leqslant k \leqslant h\}$

$r = 2$

While $L_r \neq \varnothing$ do{

　$C_{r+1} = \mathrm{fuzzy\_apriori\_gen}(L_r, \mathrm{minsup})$ //产生候选项集

　foreach $t \in T$ do{

　　$C_t = \mathrm{Subset}(C_k, t)$ //识别属于 t 的所有候选集

　　foreach $c \in C_t$ do{

　　$\mu_c(t) = \prod(\mu_{c[i]}(t))$

　　c.count = c.count + $\mu_c(t)$; //支持度计数增值

　　}

　}

　$L_{r+1} = \{c \mid \mathrm{c.count} \geqslant \mathrm{minsup}\}$ //提取频繁 $k-$项集

　$r = r + 1$

}

$L = U_k L_K$

算法 FFP:首先扫描数据库,依据专家提供的隶属函数进行数据的模糊化工作。接着计算每条记录对 1 项模糊模式的支持度累加和,然后计算在事务数据库中的支持率,从而求得频繁 1 项模糊模式 L_1,接下来,由 L_1 通过自身连接产生候选 2 项候选模糊集 C_2,依据 Apriori 性质进行剪枝,通过最小支持度 minsup 阈值筛选出 $L_2, L_3, \cdots, L_k$。最后得出所有的频繁模糊模式。由频繁模糊模式生成关联规则与传统的 Apriori 算法类似,不再叙述。

隶属函数的确定是模糊关联规则挖掘中的关键一步。对于隶属函数的确定,有统计学派与非统计学派两种不同观点与处理方法。模糊集合所表达的模糊不确定性大多是人脑对客观事物的一种直观反映。这就加剧了模糊集合隶属度的隶属函数确定的复杂性和多样性。很难用统一的模式来确定隶属函数。目前比较通用的方法是专家建议或是人们的直观感觉;而随着相关技术的发展,逐渐出现了具有自组织学习功能的模糊聚类方法和统计分析方法。还有基于三角算子的隶属函数确定方法以及采用遗传算法来提取特征函数从而确定和优化隶属函数的方法。文中引人文献[7]

中隶属函数的生成方法，该方法是一种基于聚类的自适应生成隶属函数方法。

3 算法分析与实验

提出的模糊关联规则挖掘算法是基于 Apriori 框架的，因而其计算复杂度受以下因素影响：支持度阈值、项数、事务数和事务的平均宽度。在生产 L_1 对于每个事务，需要更新事务库中出现的每个项的支持度计数。假定 w 为事务的平均宽度，则该操作所需要的时间为 $O(Nw)$，其中 N 为事务的总数。为产生候选 k - 项集，需要合并一对频繁 $(k-1)$ - 项集，确定它们是否至少有 $k-2$ 项相同。每次合并操作，最多需要 $k-2$ 次相等比较。在相等情况下，每次都产生一个可行的候选 k - 项集；在最坏的情况下，算法必须合并上次迭代发现的每对频繁 $(k-1)$ - 项集。因此，合并的总开销为：

$$\sum_{k=2}^{w}(k-2)\mid C_k\mid < \text{合并开销} < \sum_{k=2}^{w}(k-2)\mid F_{k-1}\mid^2$$

下面通过试验来证明算法的有效性，试验在 Intel Pentium 4 2.80GHz CPU，512 内存，Window XP 操作系统的 PC 机上进行。真实数据使用 UCI 机器学习数据库中的 mushroom 数据集(9124 条记录，120 个项目，平均事务强度为 23)。源代码采用 Visual C++ 6.0 实现。实验结果如图 1 和图 2 所示。

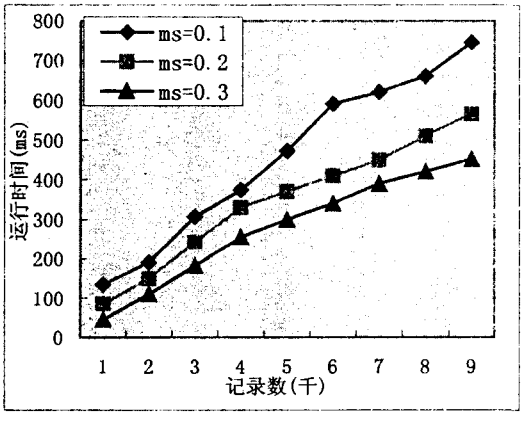

图 1 时间复杂性示意图

图 1 可以看出算法的执行时间随着记录数的增加而呈近线性增长。随着支持度阈值的减少，算法的执行时间显著增加。这是由于较小的支持度阈值会导致较多的候选集，从而增加了扫描数据库的次数，延长了程序的执行时间。由图 2 可以看出最小支持度阈值和最小置信度阈值都和规则数量成负相关性。当最小支持度阈值、最小置信度阈值都增大时，产生的规则数量会大幅减少，这和算法分析的结果一致。

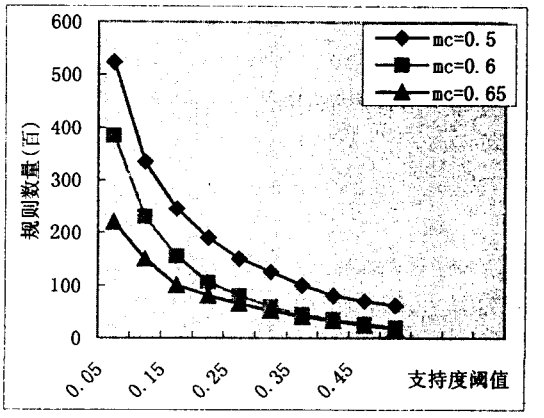

图 2 规则数量示意图

4 结论与展望

传统的关联规则挖掘算法仅限于处理非对称二元属性和分类属性。为了能够处理数量属性提出了一种宽度优先的数量关联规则挖掘算法。该算法对数据库的量化属性处理不采用区间划分法，而采用模糊概念对其进行抽象、概括，从而使得最终挖掘出的规则表示自然、易于专家理解。还改进了模糊关联规则支持度的定义，验证了提出算法的有效性。

参考文献：

[1] Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules[C]//Proceedings of the 20th International Conference on Very Large Databases (VLDB'94). Santiago: Morgan Kaufmann Publisher, 1994: 487-499.

[2] Han J, Pei J, Yin Y. Mining Frequent Patterns without Candidate Generation[C]//In Proc ACM-SOIGMOD Int. Conf. on management of data (SIGMOD'00). Dallas, TX: [s. n.], 2000: 1-12.

[3] 程继华，施鹏飞，郭建生. 模糊关联规则及挖掘算法[J]. 小型微型计算机系统，1999，20(4)：270-274.

[4] 孙建勋，陈锦云，张曙红. 用模糊方法挖掘量化关联规则[J]. 计算机工程与应用，2003(18)：190-192.

[5] 王 咏，申瑞民. 运用模糊集挖掘数量属性数据的关联规则[J]. 计算机仿真，2004，21(8)：129-131.

[6] Tan Pang-Ning, Steinbach M, Kunmar V. Introduction to Data Mining[M]. Beijing: People's post & telecommunications publishing house, 2006.

[7] 孙 逊，胡光锐，李剑萍. 一种基于模糊聚类的隶属函数定义方法[J]. 计算机应用与软件，2005，20(7)：86-88.