基于云的概念划分及其在关联采掘上的应用*

杜 鹢[1], 李德毅[2]

[1](中国电子系统工程公司网管中心,北京 100840);

[2](中国电子系统工程公司研究所,北京 100036)

E-mail: yidu@sina.com

摘要: 将数量型属性转换为布尔型属性是数量型属性关联规则采掘的主要方法,但如何使区间的划分合理一直是研究的热点.传统的划分方法由于不能反映数据间的实际分布规律或者是划分的边界过硬,使得最终都不能得到令人容易理解的关联知识.提出了一种基于云模型的新划分方法——云变换,可以有效地根据数据的实际分布将数量型属性的定义域划分为多个基于云的定性概念,这种划分摒弃了以前的硬划分,使得到的结果除了保留传统硬划分所具有的优点以外,也更加符合实际的数据分布和人的思维方式,从而最终得到概括的、易理解的、有效的关联规则.

关键词: 关联规则;云模型;云变换

中图法分类号: TP181 **文献标识码**: A

自文献[1]提出有关关联规则的采掘方法及相应算法以来,有关关联规则采掘的研究一直是数据采掘领域的研究热点[2~4],它主要是指在满足最小支持率和最小可信度的条件下,从数据库中采掘出诸如"顾客购买牛奶→同时购买面包"这样的知识.目前的研究主要集中在对布尔型关联规则的研究上,而在实际应用中,由于许多数据库中的数据都是数量型数据(如工资、年龄),因而数量型属性关联规则的研究逐渐成为研究的重点,并相应地提出了一些算法.其主要的思想都是将数量型属性的定义域通过区间划分的方法转换为布尔型属性,然后再利用布尔型关联规则的采掘算法进行挖掘.但这样会带来一个明显的问题,即数量型属性的定义域该如何划分,文献[5]提出了在区间划分中很容易出现的"catch-22"问题:

最小支持率问题:如果区间划分过小,会使包含此区间的规则的支持率很低,从而会造成规则产生的数量过少.

最小信任度问题:如果区间划分过大,会使包含此区间的规则的信任度很低,从而造成规则产生的数量过少;同时,区间划分过大,规则所包含的信息量也会相应地减少.

即使在区间的划分上用上面的两个问题进行了平衡,但如何合理、有效地划分属性区间,使其能够真实地反映此属性中数据在定义域中的实际分布则是采掘数量型属性关联规则的关键问题.

为此,许多文献提出了不同的解决方法.文献[5]提出了最为简单的方法,即将属性区间等分.但这种方法不能反映实际的数据分布,它完全靠人为的方法进行区间定义,因此得到的关联规则可能没有什么实际意义,而有意义的关联规则却可能因为区间的划分不合理而无法得到.文献[6]提出的方法是先将定义域分割成非常小的区间,然后将相邻小区间(交易数小于 $c * S_{ave}$)逐步合并成

* **收稿日期**: 1999-07-29; **修改日期**: 1999-12-03

基金项目: 国家863高科技发展计划资助项目(863-306-ZT06-07-02)

作者简介: 杜鹢(1971—),男,河北辛集人,博士,工程师,主要研究领域为数据挖掘,数据仓库,网络管理;李德毅(1941—),男,江苏泰州人,博士,研究员,博士生导师,中国工程院院士,主要研究领域为数据挖掘,智能控制,指挥自动化,系统工程.

有意义的大区间(交易数大于 $c*S_{ave}$),但由于合并方式不惟一,从而可能得到不同的划分结果,并且最终的结果仍然是一种硬划分.文献[7]提出了一种基于距离进行区间划分的方法,虽然得到的结果较好,但和文献[6]一样,仍然没有摆脱硬划分的束缚.文献[8]提出了在不减少信息丢失的情况下进行区间合并的方法,但由于这种方法涉及多个属性之间的关系,因此实际可操作性较小.

根据上面的问题,本文提出了一种基于云模型对数量型属性进行划分的概念划分算法.此方法可根据数据的实际分布将其划分为多个基于云的概念.这种划分的特点是,所得到的概念反映了此属性中数据在定义域中的实际分布,同时,由于概念的边界是模糊的,不确定的,因而是一种软划分方法,这样所得到的结果集更加符合人的思维,同时又保持了传统硬划分所具有的优点.

1 基本概念

1.1 云模型

定义 1[9]. 设 X 是一个普通集合 $X=\{x\}$,称为论域.关于论域 X 中的模糊集合 $\tilde{A}$,是指对于任意元素 x 都存在一个有稳定倾向的随机数 $\mu\tilde{A}(x)$,叫做 x 对 $\tilde{A}$ 的隶属度.隶属度在基础变量上的分布称为云.在对模糊集的处理过程中,论域中某一点到它的隶属度之间的映射是一对多的转换,而不是一条明晰的隶属曲线,从而产生了云的概念.在云模型中,经过映射,属于一个定性语言值的数值是不确定的,始终在细微变化着,并且这种变化不剧烈影响到云的整体特征.云可伸缩、无边沿、有弹性,云滴的分布特性反映了映射的模糊性和随机性,其整体形状是最重要的.

1.2 云的数字特征

由于社会和自然科学中的大量模糊概念(特别是常识性知识的表述),其期望曲线都近似服从正态或半正态分布,因而基本云即正态云是表征语言原子最重要、最有力的工具,比如青年、工资高等语言原子用云都可以很好地描述,而云的数字特征,则反映了定性知识的定量特性.更为简单、方便的是,一个基本云只需要用期望值 Ex、熵 En、超熵 He 这 3 个数字特征就可以完整地表征出来.

期望 Ex:在普通正态云的论域 X 中,对应于隶属度最大值的基础变量 x 称为云的期望,它标定了云对象在论域中的位置,即云的重心位置,换句话说,Ex 反映了相应的模糊概念的信息中心值.

熵 En:概念模糊度的度量.熵的大小直接决定了在论域中可被模糊概念所接受的范围.

超熵 He:可谓熵 En 的熵,反映了云的离散程度.超熵的大小间接地反映了云的厚度.

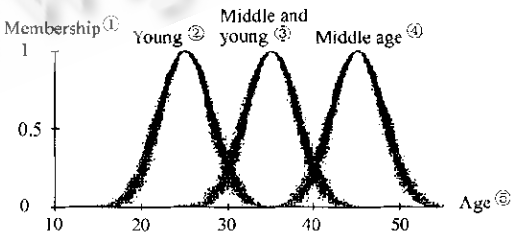


Fig. 1 Cloud model

图 1 云模型

对模糊集 $\tilde{A}$ 而言,重要的是云的形状所反映出的整体特性,以及使用时隶属度所呈现的规律性,例如,图 1 就是用云所表示的"青年"、"中青年"和"中年"这 3 个概念,其中"青年"的数字特征为 $Ex=25, En=3, He=0.2$,"中青年"的数字特征为 $Ex=35, En=3, He=0.2$,而"中年"的数字特征

分别为 $Ex=45, En=3, He=0.2$.

云模型主要有以下几个特点:

(1) 所描述的概念的数值具有凝聚性,例如,在图 1 描述"青年"的云中,25 附近的点最密,离 25 越远,点越稀.

(2) 云的期望曲线服从正态分布,便于反映大量日常的模糊概念.

(3) 对于相同的 x,其隶属于概念的隶属度具有随机性,会在一定的范围内浮动,这恰好反映了不同的人对同一事物看法的差异.

2 基于云模型的概念划分方法

2.1 云变换

很显然,利用云模型可以将数量型属性的定义域划分为多个由云模型表征的概念,但如何使划分得到的概念能够反映此属性中数据的实际分布,则是本文讨论的重点,为此,我们引入一个新概念:云变换.

定义 2. 云变换是指对于任意一个不规则的数据分布,根据某种原则进行数学变换,使之成为若干个不同的云的叠加,即 $g(x)\approx\sum_{j=1}^{m}c_j * f_j(x)\left(0<g(x)-\sum_{j=1}^{m}c_j * f_j(x)<\varepsilon\right)$,其中 $g(x)$ 为数据的分布函数,$f_j(x)$ 为基于云的概率密度期望函数,c_j 为系数,m 为叠加的云的个数,ε 为用户定义的可允许的最大误差.

根据云变换的定义可知,将数量型属性 i 的定义域划分为 m 个概念的问题可演变成对公式 $g(x)\approx\sum_{j=1}^{m}g_j(x)$ 的求解,其中 $g_j(x)$ 表示第 j 个概念的数据分布期望函数. 而对于任意一个 $g_j(x)$ 而言,有 $g_j(x)=c_j * f_j(x)$ 成立,其中 $f_j(x)$ 为第 j 个概念的概率密度期望函数,c_j 为系数,由于概念 j 是由云刻画的,因而 $f_j(x)$ 由云中的数字特征来确定,最终对数量型属性 i 的定义域的概念划分就演变为对公式 $g(x)\approx\sum_{j=1}^{m}c_j * f_j(x)$ 的求解,也就是对每一个概念的数字特征 Ex_j, En_j 和 c_j 的求解过程.

本文的最终目的,就是要将属性 i 中的数据划分为多个基于云模型的概念:使得在同一概念中的数据彼此聚集,不同概念之间的数据彼此相对分离. 下面,我们结合实例来说明概念划分的具体过程.

2.2 实 例

假设给定实验数据库表 T,其中数量型属性 i 的定义域为整型,而属性中属性值的实际取值范围为 $[0,254]$,数据库 T 中总记录数为 $n=43660$.

步骤 1. 对属性 i 定义域中的每一个可能的取值 $x(x=0,1,\ldots,254)$,计算数据库中含有该属性值的记录个数 y,得到属性 i 的数据分布函数 $g(x)(a\leqslant x\leqslant b)$(如图 2 所示).

步骤 2. 寻找数据分布函数 $g(x)$ 的波峰值所在的位置,将其属性值定义为云的重心位置(期望) $Ex_j(j=0,\ldots,m-1)$;然后计算用于拟合 $g(x)$ 的以 Ex_j 为期望的云模型的熵,计算云模型的数据分布函数 $g_j(x)$(如图 3 所示).

步骤 3. 从 $g(x)$ 中减去已知云模型的数据分布 $g_j(x)$,得到新的数据分布函数 $g'(x)$(如图 4 所示),并在此基础上重复步骤 2 和步骤 3,得到多个基于云的数据分布函数 $g_j(x)$,图 5 表示了各

个概念数据分布函数 $g_i(x)(a\leqslant x\leqslant b)$,图 6 表示了各概念的概率密度期望函数曲线 $f_j(x)$.

步骤 4. 根据已知的 $g(x)$、最后得到的误差分布函数 $g'(x)$以及上面的结果,得出基于云模型的每个概念的 3 个特征值,而相应划分得到的概念如图 7 所示.

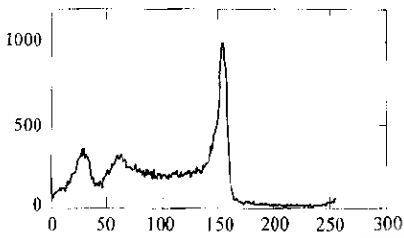


Fig. 2 Distribution of experiment data

图 2 实验数据分布图

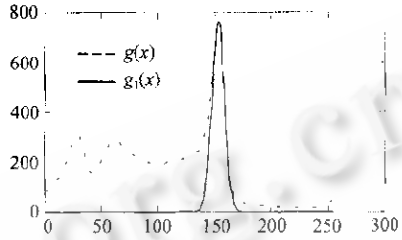


Fig. 3 Data distribution function $g_1(x)$ based on cloud model

图 3 基于云模型的数据分布函数 $g_1(x)$

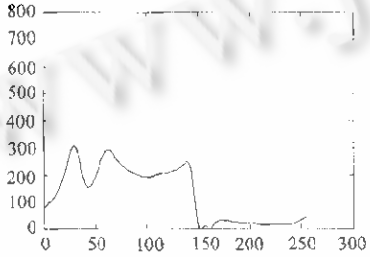


Fig. 4 Data distribution function $g'(x)$ which subtract $g_1(x)$

图 4 减去 $g_1(x)$后的数据分布函数 $g'(x)$

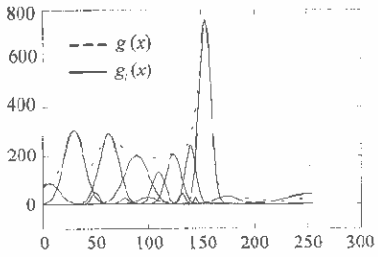


Fig. 5 Data distribution function $g_j(x)$

图 5 数据分布函数 $g_j(x)$

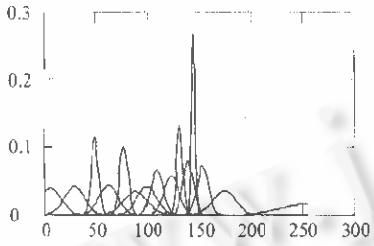


Fig. 6 Probability density expectation function $f_j(x)$

图 6 概率密度期望函数 $f_j(x)$

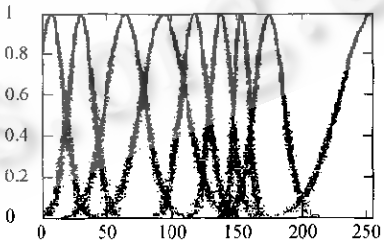


Fig. 7 The result of concept partition

图 7 概念划分结果

在图 7 中,云滴表示数据库 T 中的实际数据,x 轴表示属性 i 的定义域,y 轴表示每个云滴隶属于各自概念的隶属度.这样,通过云变换,把一个任意不规则的分布,根据某种规律进行数学变换,使之成为若干个大小不同的云的叠加,叠加的云越多,越能反映实际的数据分布,产生的误差也就越小.

2.3 概念划分算法

基于云变换的定义,本文给出了利用峰值法进行概念划分的算法.

算法 1. 概念划分算法

输入:需要进行概念划分的属性 i 的定义域,属性 i 对应的所有属性值

输出:属性 i 中的 m 个概念及其对应的数字特征

Begin

(1) 统计属性 i 的每一个可能的取值 $x(a\leqslant x\leqslant b)$ 在数据库 T 中的实际个数 y,得到属性 i 的实际数据分布函数 $g(x)$;

(2) $j=0$;

(3) $Clouds=\Phi$;$g'(x)=g(x)$;

(4) **While** $\max(g'(x))>\varepsilon$

//寻找属性 i 的数据分布函数 $g(x)$ 的波峰值 c_j 所在位置为云模型的重心位置(期望)

(5) $Ex_j=\text{Find\_}Ex(g(x))$;

//计算用于拟合 $g(x)$ 的以 Ex_j 为期望的云模型的熵

(6) $En_j=\text{Find\_}En(c_j,Ex_j,\varepsilon)$;

//计算云模型的数据分布函数

(7) $g_j(x)=c_j*Cloud(Ex_j,En_j)$;

//减去已知云模型的数据分布

(8) $g'(x)=g'(x)-g_j(x)$;

(9) $j=j+1$;

(10) **end While**

(11) **for** $j=0$ to $m-1$ **do**

(12) $Clouds(Ex_j,En_j,He_j)=\text{Calculate\_}He(g(x),g'(x),Cloud(Ex_j,En_j))$;

(13) **end for**

End

算法中在寻找合适的 En_j 时,是通过计算云模型的期望曲线 $y=e^{-\frac{(x-Ex)^2}{2(En)^2}}$ 在 $Ex_j-5*En_j\sim Ex_j+5*En_j$ 范围内与 $g'(x)$ 进行拟合,当拟合后的误差小于允许的误差范围 ε 后,即认为找到了合适的 En_j.

2.4 讨 论

2.4.1 m 的确定

通常,概念划分的个数 m 有 3 种确定方法:一种是由用户事先定义,文献[4,5]就是采用这种方法;另一种是通过人工交互对得到的概念进行重新调整;还有一种是由用户指定可允许的最大误差范围 ε,ε 越小,m 就越大,ε 越大,m 就越小.本文采用的是第 3 种方法,即由用户定义可允许的最大误差 ε.我们知道,由于数据分布一般是不规则的,而人们又希望用有规律的数学公式表示它,因此,当用有限个概念表示原有的数据分布时,它们之间就会存在误差,而 ε 表示了每一个基于云的概念与原数据分布函数进行拟合时它们之间所允许相差的最大记录个数.用户给定的误差越小,得到的概念个数 m 就会越大,而当给定的误差 ε 为 0 时,则 $m\to\infty$,这时 $g(x)=\sum_{j=1}^{m}g_j(x)$.如果用户认为给定的 ε 过大或过小,而使得到的概念个数 m 过少或过多时,用户可以利用第 2 种方法调整 ε 的大小,直至得到满意的概念个数为止.或者用户直接对已得到的概念进行概念合并或分解,以得到合适的概念个数.由于本算法在运行过程中及最终结果上都是可视化的,因而用户的操作和调整也是简单而直观的.

2.4.2 算法复杂性

本算法的时间复杂性为 $O(n)+O(m*w)$.在第 1 部分中,n 表示数据库中的记录数,$O(n)$ 为

算法中步骤1的执行时间.在步骤1中,需要将数据库中属性i的所有记录扫描一遍,从而得到属性i的数据分布函数.第2部分$O(m*w)$为算法第4~13步的执行时间.在$O(m*w)$中,w表示属性i在数据库中实际的取值范围,它代表了算法在执行过程中的扫描范围;m则代表了最终得到的概念的个数,它表达了算法从第4~10步的循环次数,每循环一次则可产生一个新的概念,通过前面的讨论可知,m的大小是由用户定义的误差的大小来决定的,误差越小,相应的循环次数也就会越多.

2.4.3 算法的特点

与模糊划分等其他方法相比,利用云变换对数量型属性进行概念划分有以下几个特点:

· 由于云模型随机性的特点,定义域中的元素对概念的隶属程度具有统计意义上的随机性,因而概念的边界是模糊不清的.这样就使得在边界上属性值相同的元素由于其隶属度的不同而可能被划分到不同的概念当中;

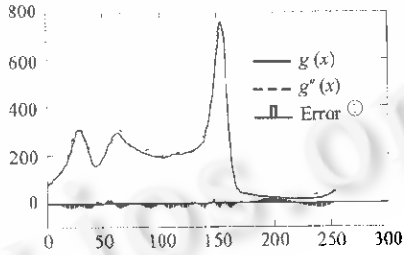
· 从概念的划分过程可以看出,出现频率高的元素对概念的贡献大于出现频率低的元素,因而划分后得到的概念能够较好地反映定义域中数据的聚集情况和实际分布情况;

· 划分的概念越多,产生的误差越小;划分的概念越少,产生的误差就越大;

· 本算法是一种动态划分方法,随着数据库中数据的不断变化,用此算法产生的概念也会发生相应的变化.

2.5 实验结果

依据算法,本文对实验数据库表T中的数量型属性i的定义域进行了概念划分,共得到15个概念,图8表示了实际数据分布函数$g(x)$、由15个概念叠加后得到的数据分布函数$g''(x)$以及两者的误差.



①误差.

Fig.8 Experimental result

图8 实验结果

3 增强型Apriori算法

在概念划分算法的基础上,本文对经典的Apriori算法进行了适当的修改,得到采掘数量型属性关联规则的增强型Apriori算法.

算法2. 增强型Apriori(数量型属性关联规则采掘算法).

(1) 利用概念划分算法将数据库中各数量型属性的定义域划分为多个概念;

(2) 将各属性值映射到不同的概念中;

(3) 从各概念中找出有价值的概念,组成强项集,通过不同属性概念之间的不断组合,得到最后的频繁项集;

(4) 应用频繁项集得到信任度超过最小信任度的关联规则;

在采掘算法的第(2)步中需将属性中的数据映射到各个概念当中去,这时可计算属性 i 中每个数值隶属于各概念的隶属度,并将其划归到隶属度最大的那个概念中去.由云的概念可知,输入同样的数值会得到不同的隶属度.因而在两个概念相交的边缘区域会出现这样的情况:记录 t_i 中属性 i 的数值与 t_j 中属性 i 的数值相同,但它们分别属于两个不同的概念.这正充分体现了云的随机性与模糊性的统一,较好地解决了硬划分带来的问题.

4 结 论

本文提出了一种基于云进行数量型属性区间划分的方法,即云变换.基于云进行概念划分的好处在于,由于云具有的模糊性与随机性,当数据库的记录数在一定的程度上增加时,概念的划分会保持相对的稳定,而不会像硬划分那样,可能会引起区间划分的突变.

本文是在基本云的基础上进行区间划分的,而在有时候,某些概念的期望不再像基本云那样只是一个数值,而是一个区间,这时,一个云就需要有 7 个数字特征来刻画(基本云实际上是这种云的特殊情况).这虽然复杂了一些,但是更能适合多种场合.利用这种云进行数量型属性的概念划分是我们今后工作的主要内容.

References:

[1] Agrawal, R., Imielinske, T., Swami, A. Mining association rules between sets of items in large databases. In: Buneman, P., Jajodia, Sushil, eds. Proceedings of the ACM SIGMOD International Conference on the Management of Data. Washington, DC: ACM Press, 1993. 207~216.

[2] Heikki, Mannila, Hannu, Toivonen, Inkeri, Verkamo A. Efficient algorithms for discovering association rules. In: Fayyad, Usama M. Uthurusamy Ramasamy, ed. Workshop on Knowledge Discovery in Databases. Seattle, Washington: AAAI Press, 1994. 181~192.

[3] Park, J. S., Chen, M. S., Yu, P. S. An effective hash-based algorithm for mining association rules. In: Carey, M. J., Schneider, D. A., eds. Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data. San Jose, CA: ACM Press, 1995. 175~186.

[4] Savasere, A., Omiecinski, E., Navathe, S. An efficient algorithm for mining association rules in large databases. In: Dayal, Umeshwar, Gray, P. M. D., Nishio, Shojiro, eds. Proceedings of the 21th International Conference on Very Large Data Bases. Zürich, Switzerland: Morgan Kaufmann Publishers, 1995. 432~444.

[5] Srikant, R., Agrawal, R. Mining quantitative association rules in large relational tables. In: Jagadish, H. V., Mumick, Inderpal Singh, eds. Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data. Montreal, Canada: ACM Press, 1996. 1~12.

[6] Zhang, Zhao-hui, Lu, Yu-chang, Zhang, Bo. An algorithm for mining quantitative association rules. Journal of Software, 1998,9(11):801~805 (in Chinese).

[7] Miller, R. J., Yang, Y. Association rules over interval data. In: Peckham, J., ed. Proceedings of the ACM SIGMOD International Conference on Management of Data. Tucson, Arizona: ACM Press, 1997. 452~461.

[8] Wang, Ke, Soon, Hock William Tay, Liu Bing. Interestingness Based interval merger for numeric association rules. In: Agrawal, R., Stolotz, P., eds. Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining. New York: AAAI Press, 1998. 121~128.

[9] Li, De-yi, Meng, Han-jun, Shi, Xue-mei. Cloud model and cloud model generator. Computer Research and Development, 1995,32(6):15~20 (in Chinese).

附中文参考文献:

[6] 张朝晖,陆玉昌,张钹.发掘多值属性的关联规则.软件学报,1998,9(11):801~805.

[9] 李德毅,孟海军,史雪梅.隶属云和隶属云发生器.计算机研究与发展,1995,32(6):15~20.

Concept Partition Based on Cloud and Its Application to Mining Association Rules*

DU Yi[1], LI De-yi[2]

[1](*Network Management Center, China Electronic System Engineering Company, Beijing* 100840, *China*);
[2](*Institute of China Eletronic System Engineering Company, Beijing* 100036, *China*)
E-mail: yidu@sina.com

Abstract: Converting quantitative attributes into Boolean attributes is the general way for mining quantitative association rules. So how to reasonably partition domain values is very important. Traditional method can not get the easy to understand knowledge because it can not reflect the actual data distribution or the partition is too sharp. In this paper, a new method——cloud transform, which uses many concepts represented by cloud model to fit the real distribution of data—— is introduced. This method can reflect the distribution of data in that domain while keeping the soft boundaries. Therefore, the discovered association rules are also easy to understand.

Key words: association rule; cloud model; cloud transform

* Received July 29, 1999; accepted December 3, 1999
Supported by the National High Technology Development Program of China under Grant No. 863-306-ZT06-07-02