挖掘语言值关联规则*

陆建江[1]，宋自林[2]，钱祖平[1,3]

[1](解放军理工大学 通信工程学院,江苏 南京 210016);

[2](解放军理工大学 计算机与指挥自动化学院,江苏 南京 210016);

[3](东南大学 微波与毫米波国家重点实验室,江苏 南京 210096)

E-mail: ljj666@sina.com

摘要: 讨论了大型数据库上数量属性的关联规则问题.为了软化论域的划分边界,应用相关的模糊c-方法(relational fuzzy c-means,简称RFCM)算法确定正态模糊数的两个参数,并借助正态模糊数模型来划分数量属性的论域,由此生成一系列的语言值关联规则.另外,给出了语言值关联规则的挖掘方法.由于语言值能很好地表示抽象的概念,从而使得挖掘出的关联规则更抽象、更容易被人理解.

关键词: 数据挖掘;相关的模糊c-方法算法;正态模糊数;语言值;关联规则

中图法分类号: TP311 **文献标识码**: A

在信息处理领域,数据挖掘是数据库研究的一个新课题,而关联规则的挖掘在数据挖掘的研究中更是受到许多人的重视.文献[1]首先提出挖掘关联规则的思想,讨论了布尔型属性的关联规则的挖掘问题.文献[2]讨论了数量属性的关联规则问题.文中的算法将数量属性的论域划分成多个区间,从而把数量属性的关联规则问题转换成布尔型关联规则的问题来进行讨论.一种方法是将属性的论域划分成离散的互不重叠的区间,由于明显的区间划分会将区间附近的一些元素排斥在外,从而导致一些有意义的区间可能被忽略.例如,考虑一个城市的降水量,对于通过划分得到的一个区间[20,40],由于划分的边界过硬,可能会使得这个区间的支持率小于用户给定的最小支持率,但如果将41,42的记录考虑进来的话,此时区间[18,42]就可能变得有意义.另一种方法是将属性的论域划分成重叠的区间,这时,处于边界附近的元素就可能同时处于两个区间,因而会过分强调这些元素的作用.本文通过一个实例阐述大型关系数据库中数量属性的语言值关联规则的挖掘方法.第1节应用RFCM(relational fuzzy c means)算法确定正态模糊数的两个参数,并借助正态模糊数模型来软化论域的划分边界.第2节组合第1节的正态模糊数,由此生成一系列的语言值关联规则,接着利用给定的阈值从大型数据库中挖掘出所有有意义的语言值关联规则.第3节给出结论.

1 软化划分边界

文中实例的数据来自文献[3],数据库记录了1986、1987、1989年30个城市的日照时数(小时/每月)、降水量(毫米/每月)、每月平均气温(度),共有1 080条记录.本节应用RFCM算法将日照时数、降水量、每月平均气温的1 080个取值划分成7个语言值:极大、很大、大、一般、小、很小、极

* 收稿日期: 2000-02-15; 修改日期: 2000-05-30

基金项目: 国家自然科学基金资助项目(69931040)

作者简介: 陆建江(1968-),男,江苏人,博士生,讲师,主要研究领域为数据挖掘,数据仓库,模糊数学;宋自林(1944-),男,安徽人,教授,博士生导师,主要研究领域为军事通信学,数据仓库,数据挖掘;钱祖平(1961-),男,江苏人,博士,副教授,主要研究领域为微波理论与技术,电磁成像,微分方程的数值计算方法,数据挖掘.

© 中国科学院软件研究所 http://www.jos.org.cn

小,并将这些语言值表示成正态模糊数模型:$y=\exp\left[\frac{(x-\mu)^2}{\sigma^2}\right]$. 设目标数据集 $X=\{x_1,\ldots,x_n\}\subset R^p$,

$$M_{fc}=\{U\in R^{cn}\mid 0\leqslant u_{ik}\leqslant 1,\ u_{1k}+u_{2k}+\ldots+u_{ck}=1 \text{ for } 1\leqslant k\leqslant n,$$
$$\text{and } u_{i1}+u_{i2}+\ldots+u_{in}>0 \text{ for } 1\leqslant i\leqslant c\}.$$

把 X 划分成与 c 类相关的模糊 c-方法(RFCM)算法[4]如下:

(RFCM-1) 计算相关数据 $R=[r_{ij}]$,这里,r_{ij}是目标数据 x_i 到 x_j 的距离的平方. 取定 c,$2\leqslant c<n$,取定 $m>1$,初始化矩阵 $U^{(0)}\in M_{fc}$,设置循环次数 s,$s=0,1,2,\ldots$;

(RFCM-2) 用 $U=U^{(s)}$ 计算 c 个向量 $v_i=v_i^s$,$v_i=(u_{i1}^m,u_{i2}^m,\ldots,u_{in}^m)^T\Big/\sum_{k=1}^{n}(u_{ik})^m$;

(RFCM-3) 修改 $U=U^{(s+1)}\in M_{fc}$. 方法如下:记 $(d_{ik})^2=(Rv_i)_k-(v_i^TRv_i)/2$,如果对每个 i,$d_{ik}>0$,则 $u_{ik}=1\Big/\sum_{j=1}^{c}(d_{ik}/d_{jk})^{2/(m-1)}$;否则,如果 $d_{ik}>0$,则 $u_{ik}=0$,$u_{ik}\in[0,1]$,$u_{1k}+u_{2k}+\ldots+u_{ck}=1$;

(RFCM-4) 取矩阵范数 $\|\ \|$,如果 $\|U^{(s+1)}-U^{(s)}\|\leqslant\varepsilon$,则循环停止;否则,置 $s\leftarrow s+1$ 并返回(RFCM-2).

例如,划分日照时数成 7 个语言值的过程如下:取 $r_{ij}=|x_i-x_j|^2$,$m=2$,$\varepsilon=0.001$,$c=7$,矩阵范数 $\|\ \|$ 为矩阵中元素的最大值,初始化矩阵 $U^{(0)}\in M_{fc}$,$U^{(0)}$中的元素不全相等. 用 RFCM 算法进行聚类,最后得到划分矩阵 U 和 7 个 v_j,比较 v_i,最大的 v_i 所对应的 U 中行的元素即是语言值极大对 1 080 个样本点的隶属度,根据这些隶属度可以确定表示语言值极大的正态模糊数的两个参数. 设语言值极大的中心为 v,1 080 个样本点的隶属度为 r_i,$i=1,2,\ldots,1\,080$. 正态模糊数的参数 μ 取为中心 v. 为了求参数 σ,可采用正态模糊数 $y=\exp\left[-\frac{(x-\mu)^2}{\sigma^2}\right]$ 去逼近 1 080 个样本点的隶属度所构成的曲线,即求满足目标函数 $g(\sigma)=\sum_{i=1}^{1080}\left[\exp\frac{-(x_i-\mu)^2}{\sigma^2}-r_i\right]^2$ 的最小的 σ. 为了避免用迭代法求解非线性方程,可转化为求满足目标函数 $h(\sigma)=\sum_{i=1}^{1080}\left[-\frac{(x_i-\mu)^2}{\sigma^2}-\ln r_i\right]^2$ 的最小的 σ. 令 $\frac{1}{\sigma^2}=t$,$-(x_i-\mu)^2=a_i$,求出 $t=\sum_{i=1}^{1080}a_i\ln r_i\Big/\sum_{i=1}^{1080}a_i^2$. 表 1 列出了日照时数的 7 个语言值的正态模糊数的参数,其他属性可作同样处理. 由于正态模糊数是一条光滑的曲线,它能在集合元素和非集合元素之间提供平滑的变迁,因此正态模糊数模型能较好地实现连续量和离散量之间的转换,并最终软化了属性论域的划分边界. 通过用正态模糊数模型来软化属性论域的划分边界,可以在挖掘规则时充分地、合理地考虑各个元素所作的贡献,从而避免了文献[2]中忽略一些元素或过分强调一些元素的缺点. 同时,由于语言值能很好地表达抽象的概念,因此使得挖掘出的语言值关联规则更贴近人的思维方式.

Table 1 Hours of sunlight (hours/month)

表 1 日照时数(小时/月)

Linguistic value①	μ	σ^2
Extremely large②	287.62	737.67
Very large③	249.19	328.7
Large④	213.55	90.06
Middle⑤	177.97	424.36
Small⑥	143.52	85
Very small⑦	106.93	324
Extremely small⑧	60.07	841

①语言值,②极大,③很大,④大,⑤一般,⑥小,⑦很小,⑧极小.

© 中国科学院软件研究所 http://www.jos.org.cn

2 挖掘语言值关联规则

设 $T=\{t_1,\ldots,t_n\}$是一个数据库,t_j 表示 T 的第 j 个元组,$I=\{i_1,\ldots,i_m\}$表示属性集,属性 i_k 的论域为 V_{i_k},$t_j[i_k]$表示属性 i_k 在第 j 个记录上的取值. $X=\{x_1,\ldots,x_p\}$,$Y=\{y_1,\ldots,y_q\}$是 I 的子集,且 $X\cap Y=\varnothing$,$V_{x_1},\ldots,V_{x_p}$和 $V_{y_1},\ldots,V_{y_p}$上分别有语言值 $f_{x_h}(h=1,2,\ldots,p)$和 $f_{y_l}(l=1,2,\ldots,q)$,记 $A=\{f_{x_h}\}$,$B=\{f_{y_l}\}$,$Z=X\cup Y$. 所要讨论的语言值关联规则为"如果 X 是 A 则 Y 是 B". 判断一个语言值关联规则是否被采用需要用到支持率和信任度,当支持率和信任度分别不小于给定的最小支持率和最小信任度时,则认为语言值关联规则被采用,否则不被采用. 下面,我们给出支持率和信任度的定义.

定义 1. 语言值关联规则"如果 X 是 A 则 Y 是 B"的支持率记为 S,其中

$$S=\frac{\sum_{t_j\in T}\left[\prod_{z_h\in Z}f_{Z_h}[t_j(z_h)]\right]}{n},\ n\text{ 是 }T\text{ 的元组个数.}$$

定义 2. 语言值关联规则"如果 X 是 A 则 Y 是 B"的信任度记为 C,其中

$$C=\frac{S}{\frac{1}{n}\sum_{t_j\in T}\prod_{h=1}^{p}f_{x_h}[t_j(x_h)]},\ n\text{ 是 }T\text{ 的元组个数.}$$

随着语言值的改变,语言值关联规则也在改变,因此所要讨论的规则是无穷的. 在实际应用中,每个属性所取的语言值不是很多,例如,在上节的实例中,可讨论语言值关联规则:"如果 X 是 A,则 Y 是 B",其中 $X=\{$日照时数,降水量$\}$,$Y=\{$每月平均气温$\}$,$A=\{f_1(x),f_2(x)\}$,$B=\{f_3(x)\}$. $f_1(x)$,$f_2(x)$,$f_3(x)$分别是日照时数、降水量、每月平均气温的语言值. 组合 7 个语言值,得到规则 343 条,这就是说,实际应用中所要讨论的规则可以认为是有限的. 为了表示简单,采用 ijk 表示规则,i 表示取日照时数的第 i 个语言值;j 表示取降水量的第 j 个语言值;k 表示取每月平均气温的第 k 个语言值. 给定最小支持率 $S=0.013$ 和最小信任度 $C=0.35$,表 2 列出了挖掘出的 10 条有意义的规则以及这些规则的支持率和信任度. 表中列出的是一些支持率和信任度大的规则,这些规则揭示了大型数据库中所包含的有意义的信息. 例如,表中第 1 行表示日照时数极大且降水量很小则每月平均气温很大这条规则被支持的程度是 0.015 063,被信任的程度是 0.400 270. 随着数量属性的取值被划分成更多的语言值,文中的方法可以挖掘出语义更具体的规则. 当然,挖掘规则所需的时间可能会增加.

Table 2

表 2

Linguistic value association rules①	Support②	Confidence③
162	0.015 063	0.400 270
163	0.013 286	0.353 075
262	0.015 043	0.359 644
275	0.015523	0.363715
375	0.016 931	0.449 526
376	0.018 640	0.494 888
476	0.041 502	0.525 749
477	0.030 686	0.388 733
576	0.017 291	0.605 135
665	0.013 876	0.370 245

①语言值关联规则,②支持率,③信任度.

© 中国科学院软件研究所 http://www.jos.org.cn

给定最小支持率$S=0.013$和最小信任度$C=0.35$,表3列出了当数量属性的取值分别被划分成5～15个语言值时在PIII 400机上挖掘规则所需的时间,从表3中可以看出,随着语言值个数的增加,挖掘规则所需的时间总体上是增加的,但当数量属性的取值被划分成5或13个语言值时,挖掘规则所需的时间出现了反常,这是因为RFCM算法计算的时间不一定随着语言值个数的增加而增加.例如,当语言值个数为13时,RFCM算法计算的时间为73秒,而当语言值个数为14个时,RFCM算法计算的时间为34秒.在实际应用中,每个属性所取的语言值一般不很多,取15个语言值已足够表达人们的思想,因此不必担忧挖掘规则所需的时间会增加到无法容忍的地步.

Table 3

表3

Linguistic value numbers①	5	6	7	8	9	10	11	12	13	14	15
Costing time (sec.)②	18	7	10	16	26	28	36	37	82	46	68

①语言值个数,②花费时间(秒).

3 结 论

本文借助正态模糊数模型软化数量属性的划分边界,生成了一系列的语言值关联规则,接着给出了规则的挖掘方法.尽管文中的挖掘方法是通过一个实例得到阐述,但这个方法很容易推广到一般的情况.由于正态模糊数可在集合元素和非集合元素之间提供平滑的变迁,因此可以在挖掘规则时充分地、合理地考虑各个元素所作的贡献;同时,正态模糊数模型表示的语言值能很好地表达抽象的概念,因此挖掘出的规则更抽象、更符合人类的思维方式.随着数量属性被划分成更多的语言值,文中的方法能挖掘出语义更具体的语言值关联规则,但挖掘规则所需的时间总体上会增加.

References:

[1] Agrawal, R., Imieliski, T., Swami, A. Mining association rules between sets of items in large databases. ACM SIGMOD Issues, 1993,22(2):207～216.

[2] Srikant, R., Agrawal, R. Mining quantitative association rules in large relational tables. ACM SIGMOD Issues, 1996,25(2):1～12.

[3] The Statistics Bureau of China. The Statistics Annual of China. Beijing: the Statistics Press of China, 1987 (in Chinese).

[4] Hathaway, R.J., Davenport, J.W., Bezdek, J.C. Relational dual of the c-means algorithms. Pattern Recognition, 1989, 22(2):205～212.

附中文参考文献:

[3] 中国统计局.中国统计年鉴.北京:中国统计出版社,1987.

Mining Linguistic Value Association Rules*

LU Jian-jiang[1], SONG Zi-lin[2], QIAN Zu-ping[1,3]

[1](*Institute of Communications Engineering, PLA University of Science and Technology, Nanjing* 210016, *China*);
[2](*Institute of Computer Science and Command Automation, PLA University of Science and Technology, Nanjing* 210016, *China*);
[3](*State Key Laboratory of Microwave and Millimeter Wave, Southeast University, Nanjing* 210096, *China*)
E-mail: ljj666@sina.com

Abstract: The issue of quantitative association rules in large databases is discussed in this paper. In order to soften partition boundary of the domain, the relational fuzzy c-means algorithm is adopted to determine two parameters of normal fuzzy numbers, then the normal fuzzy number model is adopted to partition the domain of the quantitative attributes and a series of linguistic value association rules are generated. The mining method of the linguistic value association rules is also provided. Because the abstract concepts can be well expressed with the

© 中国科学院软件研究所 http://www.jos.org.cn

linguistic values, the mined association rules are more abstract and easy to understand.

Key words: data mining; relational fuzzy c-means algorithm; normal fuzzy number; linguistic value; association rule

* Received February 15, 2000; accepted May 30, 2000
Supported by the National Natural Science Foundation of China under Grant No. 69931040

© 中国科学院软件研究所 http://www.jos.org.cn