加权模糊关联规则的研究*)

陆建江
(解放军理工大学通信工程学院 南京210007)

Research on Weighted Fuzzy Association Rules

LU Jian-Jiang
(PLA University of SCI.&TECH,Nanjing 210007)

Abstract Algorithms for mining quantitative association rules consider each attribute equally, but the attributes usually have different importance. Two kinds of algorithms for mining the weighted fuzzy association rules are provided with respect to two kinds of database. The first algorithm can effectively consider the importance of quantitative attributes, and considers that the importance of association rule is not increased with the amount of attributes in the rule. The second algorithm not only considers the importance of quantitative attributes, but also considers that the importance of association rule is increased with the amount of attributes in the rule.

Keywords Data mining, Quantitative attributes, Association rules, Fuzzy, Support, Confidence, Fuzzy c-means

1 引言

关联规则是展示属性-值频繁地在给定的数据集中一起出现的条件,最常见的是对大型超市的事务数据库进行货篮分析,文[1]提出了解决此类问题的布尔型属性关联规则的Apriori算法。数量关联在股市分析、银行存款分析和医疗诊断等众多方面都有重要应用价值。数量关联用来描述数量型属性特征之间的相互关系,用数量型关联规则来表示,如"10%年龄在50-70之间的已婚人员至少拥有两辆汽车"。文[2]首先讨论数量型关联规则,文中的挖掘算法将数量型属性划分成多个区间,但这样的方法会引起划分边界过硬的缺点。文[3]用模糊集软化划分边界,并提出模糊关联规则的概念。文[4]用相关的模糊c-均值算法划分数量型属性,并给出语言值关联规则的挖掘算法。文[5]用正态云软化划分边界,并提出正态云关联规则的概念。文[2~5]中的挖掘算法平等一致地处理各个数量型属性,然而,数据库中的不同数量型属性往往具有不同的重要性。这类数据库包括两种类型,因此文中相应地提出了两种加权模糊关联规则的挖掘算法。第一种挖掘算法能有效地考虑数量型属性的权重,并且认为规则的重要性不随着规则中所含属性数量的增加而增加。第二种挖掘算法不仅需要考虑属性的权重,并且认为规则的重要性随着规则中所含属性数量的增加而增加。

2 数量型属性的离散化

设 $T=\{t_1,\cdots,t_n\}$ 是一个数据库,t_j 表示 T 的第 j 个记录,$I=\{i_1,\cdots,i_m\}$ 表示属性集,属性集中的属性可以是数量型属性、布尔型属性和类别型属性,第 j 个记录在属性 i_k 上的取值为 $t_j[i_k]$,$W=\{w_{i_1},\cdots,w_{i_m}\}$ 是属性的权重集,其中 w_{i_j} 为属性 i_j 的权重,表示属性 i_j 的重要程度。这节讨论如何将记录在属性上的取值划分成若干个模糊集。

设布尔型属性的两个取值为 A_1 与 A_2,则可以被划分成两个模糊集,仍记为 A_1 与 A_2,其中模糊集 A_1 定义为:当取值为 A_1 时为1,取值为 A_2 时为0,模糊集 A_2 的定义类似;类别型属性只有少量的几个取值,也可以采用与布尔型属性类似的方法划分。如果是数量型属性,则采用模糊 c-均值(fuzzy c-means,简称 FCM)算法划分。记

$$M_{fc}=\{U\in R^{cn}\mid 0\leqslant u_{ij}\leqslant 1,u_{1j}+u_{2j}+\cdots+u_{cj}=1,1\leqslant j\leqslant n,\ 1\leqslant i\leqslant c\},$$

R^{cn} 是实数域上 c 行 n 列矩阵组成的集合,X 是样本点集合,将 X 划分成 c 个模糊集的 FCM 算法[6]的迭代过程如下:

1 取定 c,$2\leqslant c<n$,取定 m,初始化矩阵 $U^{(0)}\in M_{fc}$,设置循环次数 s,$s=0,1,2,\cdots$;

2 用 $U=U^{(s)}$ 计算 c 个向量

$$v_i^s=v_i,v_i=\sum_{j=1}^{n}(u_{ij})^m x_j/\sum_{j=1}^{n}(u_{ij})^m;$$

3 修改 $U=U^{(s+1)}\in M_{fc}$。记 $d_{ij}=d(x_j,v_i)$,如果对每个固定的 i,对所有的 $1\leqslant j\leqslant n$,都有 $d_{ij}>0$,则 $u_{ij}=1/\sum_{k=1}^{c}(d_{ij}/d_{kj})^{2/(m-1)}$;否则,如果 $d_{ij}=0$,则 $u_{ij}=1$,其它都为零;

4 取合适的矩阵范数$\|\|$,取定 ε,ε 是任意小的实数,如果 $\|U^{(s+1)}-U^{(s)}\|\leqslant\varepsilon$,则停止;否则,$s=s+1$ 并返回第2步。

用 FCM 算法将记录在数量型属性 i_k 上的取值划分成 c 个模糊集的方法如下:

将记录在属性 i_k 上的取值放在一起作为样本点集合 $X=\{x_1,\cdots,x_n\}$,取 $m=2$,$d_{ij}=|x_j-v_i|$,矩阵范数$\|\|$为矩阵中元素的最大值,初始化矩阵 $U^{(0)}\in M_{fc}$,为了使得 FCM 算法能得到理想的结果,$U^{(0)}$ 中的元素应尽可能不相等。用 FCM 算法对 X 进行模糊聚类,最后得到划分矩阵 U 和 c 个中心 $v=\{v_1,\cdots,v_c\}$,根据中心的大小依此确定模糊集等级,最大的中心对应最大模糊集等级,其它类似。同时最大中心所对应的 U 中行的元素即是样本点在最大模糊集等级上的隶属度。在模糊数学中,模糊集常采用三角模糊数、正态模糊数或梯形模糊数表示,这里采用三角模糊数的表示方法。记 $\mu_k(x_i)$ 是样本点 x_i 在第 k 个模糊集上的隶属度,三角模糊数的表示方法如下:

*)得到国家自然科学基金重点项目(编号 69931040)资助。陆建江 博士,副教授,主要研究领域为数据挖掘、数据仓库、自然语言处理和模糊数学。

设 $X^k=\{x_i:\mu_k(x_i)\geqslant\mu_j(x_i),\forall j\in\{1,\cdots,c\}\}$，在 $X^k\cup\{v_k\}$ 中找出位于类中心 v_k 两侧的隶属度最小的样本点，设左侧隶属度最小的样本点为 x^l，隶属度为 $\mu_k(x^l)$，右侧隶属度最小的样本点为 x^r，隶属度为 $\mu_k(x^r)$，则第 k 个模糊集对应的三角模糊数(a,v_k,b)为：

$$f(x)=\begin{cases}\dfrac{x-a}{v_k-a} & a\leqslant x\leqslant v_k\\ \dfrac{b-x}{b-v_k} & v_k<x\leqslant b\\ 0 & x<a\text{ 或 }x>b\end{cases}$$

$$\begin{cases}a=x^l-\dfrac{\mu_k(x^l)(v_k-x^l)}{1-\mu_k(x^l)}\\ b=x^r+\dfrac{\mu_k(x^r)(x^r-v_k)}{1-\mu_k(x^r)}\end{cases}$$

3 加权模糊关联规则的第一种挖掘算法

这节讨论的数据库具有以下特点：不同数量型属性具有不同的重要性，同时规则的重要性不随着规则中所含属性数量的增加而增加。例如肿瘤诊断数据库，数据库取自 UCI Machine Learning Repository，有记录数569条，属性个数为32个，实验时取属性2到属性12：diagnosis、radius、texture、perimeter、area、smoothness 、compactness、concavity、concave points、symmetry、fractal dimension，共11个属性，分别记为 $i_0,\cdots,i_{10}$。diagnosis 为布尔型属性，有两个取值，"M"表示诊断为恶性，"B"表示诊断为良性，其它10个为数量型属性。如果关注10个数量型属性对肿瘤诊断结果的影响，并设10个数量型属性具有不同的重要性，实验时权重分别取为0.8，0.4，0.8，0.8，0.3，0.4，0.6，0.6，0.7，0.2，权重集一般由专家给出，这里的权重仅作实验数据。

首先通过数据库 $T=\{t_1,\cdots,t_n\}$ 构造一个加权数据库，加权数据库以划分得到的模糊集作为数据库的属性，由于属性是模糊集，称这些属性为模糊属性。模糊属性具有不同的权重，对应的数量型属性的权重即是这些模糊属性的权重。加权数据库中记录在模糊属性上的取值由下面的方法得到。设 i_k^l 是记录在属性 i_k 上的取值通过划分得到的一个模糊集，则模糊集 i_k^l 是加权数据库中的模糊属性，第 j 个记录在模糊属性 i_k^l 上的取值为：原数据库中第 j 个记录在属性 i_k 上的取值在 i_k^l 上的隶属度乘上属性 i_k 的权重，即 $i_k^l[t_j(i_k)]\times w_{i_k}$。在加权数据库中，不妨仍记所有模糊属性组成的集合为 I，所有记录组成的集合为 $T=\{t_1,\cdots,t_n\}$，第 j 个记录在模糊属性 y_k 上的取值为 $t_j(y_k)$，容易知道记录在模糊属性上的取值都属于区间[0,1]。

对肿瘤诊断数据库，布尔型属性 diagnosis 有两个取值："M"与"B"，可以被划分成二个模糊集：M 与 B，模糊集 M 定义为：当取值为 M 时为1，取值为 B 时为0；模糊集 B 的定义类似。其它10个数量型属性均被划分为三个用三角模糊数表示的模糊集，模糊集表示的语义为：大、中、小。这些模糊集都是加权数据库的模糊属性。表1列出了 radius、texture、perimeter 属性的三角模糊数参数，表2列出了加权数据库中前三条记录在一些模糊属性上的取值。

设 $X=\{y_1,\cdots,y_p\}$，$Y=\{y_{p+1},\cdots,y_{p+q}\}$ 是 I 的子集，$X\cap Y=\varnothing$，讨论的关联规则为"$X\Rightarrow Y$"，由于 X 和 Y 中的元素都是模糊属性，而且模糊属性具有权重，此时称关联规则"$X\Rightarrow Y$"为加权模糊关联规则。在下面的加权模糊关联规则的挖掘算法中，类似引用布尔型关联规则挖掘算法中的概念，只是在原有概念之前加上"加权模糊"。

表1 前三个数量型属性的三角模糊数参数

	大	中	小
radius	(14.76, 19.89, 42.08)	(11.16, 14.32, 19.42)	(−2.95, 11.08, 14.28)
texture	(20.41, 26.15, 56.60)	(15.22, 20.02, 25.80)	(−4.63, 15.03, 19.84)
perimeter	(97.05, 131.84, 283.95)	(72.00, 93.91, 128.62)	(−24.48, 71.58, 93.42)

表2 加权数据库中的部分值

(radius: 大)	(radius: 中)	(radius: 小)	(texture: 大)	(texture: 中)	(texture: 小)
0.504155	0.224418	0	0	0	0.305316
0.775338	0	0	0	0.212335	0.172366
0.769431	0	0	0.058781	0.314991	0

定义1 模糊属性集 $X=\{y_1,\cdots,y_p\}$ 的加权模糊支持率定义为 $WFSup(X)(\sum_{j=1}^{n}\prod_{i=1}^{p}t_j(y_i))/n$。如果 $WFSup(X)$ 不小于用户给定的加权模糊最小支持率，则称 X 为加权模糊大属性集。

定义2 加权模糊关联规则 $X\Rightarrow Y$ 的加权模糊支持率定义为 $WFSup(\sum_{j=1}^{n}\prod_{i=1}^{p+q}t_j(y_i))/n$。

定义3 加权模糊关联规则 $X\Rightarrow Y$ 的加权模糊信任度定义为 $WFconf=(\sum_{j=1}^{n}\prod_{i=1}^{p+q}t_j(y_i))/\sum_{j=1}^{n}\prod_{i=1}^{p}t_j(y_i)$。

表3 有意义的加权模糊关联规则

加权模糊关联规则	加权模糊支持率	加权模糊信任度
(radius:大)→(diagnosis: M)	0.143577	0.969482
(perimeter:大)→(diagnosis: M)	0.142135	0.976967
(area:大)→(diagnosis: M)	0.120537	0.987546
(concavity:大)→(diagnosis: M)	0.067884	0.935067
(concave points:大)→(diagnosis: M)	0.059056	0.997564
(radius:大)且(perimeter:大)→(diagnosis: M)	0.08867	0.989988
(radius:大)且(area:大)→(diagnosis: M)	0.080357	0.993243
(radius:大)且(symmetry:中)→(diagnosis: M)	0.047553	0.971217
(perimeter:大)且(area:大)→(diagnosis: M)	0.079444	0.994381
(perimeter:大)且(symmetry:中)→(diagnosis: M)	0.046804	0.978123
(radius:大)且(perimeter:大)且(area:大)→(diagnosis: M)	0.055316	0.99681

从定义2中容易看出：规则的重要性不随着规则中含有模糊属性数量的增加而增加，同时由于加权数据库中的元素已经加权，因此加权模糊支持率和加权模糊信任度的定义合理地考虑了模糊属性的权重。挖掘布尔型属性关联规则 Apriori 算法的核心之处是：如果一个属性集是大属性集，那么它的子集也都是大属性集。在加权数据库中，$t_j(i_k)$ 是区间[0,1]上的一个数，由定义1容易知道加权模糊大属性集的所有子集也都是加权模糊大属性集。因此，可以直接改进 Apriori 算法得到

加权模糊关联规则的第一种挖掘算法。

这节关注10个数量型属性对肿瘤诊断结果的影响，所以仅挖掘后件为恶性的加权模糊关联规则。给定加权模糊最小支持率0.045和加权模糊最小信任度0.93，挖掘得到11条有意义的加权模糊关联规则，表3列出了这些规则和它们的加权模糊支持率和加权模糊信任度，所谓有意义的加权模糊关联规则即是加权模糊支持率和加权模糊信任度分别不小于加权模糊最小支持率和加权模糊最小信任度的规则，这些规则揭示了数据库中所含的信息和一般规律，可以作为对新的肿瘤病例诊断的参考依据。11条有意义的规则中包含的数量型属性是一些权重大的属性，即 radius、perimeter、area、concavity 等，这表明算法能有效地考虑数量型属性的权重。再有，11条有意义的规则中最长的规则只包含三个模糊属性，这表明规则的重要性不随着规则中所含模糊属性数量的增加而增加。

4 加权模糊关联规则的第二种挖掘算法

第一种加权模糊关联规则的挖掘算法能有效地考虑数量型属性的权重，并且认为规则的重要性不随着规则中所含属性数量的增加而增加。但在有些数据库中，规则的重要性会随着规则中所含属性数量的增加而增加。例如为了追求最大利润，商场经理会优先考虑含有属性数量多的关联规则，因为这种关联规则往往能带动较多商品的促销。下面提出的第二种加权模糊关联规则的挖掘算法能有效地处理这类问题。

与上节的方法类似，首先通过数据库 $T=\{t_1,\cdots,t_n\}$ 构造一个新数据库，新数据库以划分得到的模糊集作为数据库的模糊属性。新数据库中记录在模糊属性上的取值由下面的方法得到。设 i_k^l 是记录在属性 i_k 上的取值通过划分得到的一个模糊集，第 j 个记录在模糊属性 i_k^l 上的取值为：原数据库中第 j 个记录在属性 i_k 上的取值在 i_k^l 上的隶属度，即 $i_k^l[t_j(i_k)]$。仍记所有模糊属性组成的集合为 I，所有记录组成的集合为 $T=\{t_1,\cdots,t_n\}$，第 j 个记录在模糊属性 y_k 上的取值为 $t_j(y_k)$，容易知道 $t_j(y_k)$ 属于区间[0,1]。设 $X=\{y_1,\cdots,y_p\}$，$Y=\{y_{p+1},\cdots,y_{p+q}\}$ 是 I 的子集，$X\cap Y=\varnothing$，讨论的加权模糊关联规则为"$X\Rightarrow Y$"。

定义4 模糊属性集 $X=\{y_1,\cdots,y_p\}\subset I$ 的加权模糊支持率定义为 $WFSup(X)=\sum_{i=1}^{p}w_{y_i}\times(\sum_{j=1}^{n}\prod_{i=1}^{p}t_j(y_i))/n$，如果 $WFSup(X)$ 不小于用户给定的加权模糊最小支持率，则称 X 为加权模糊大属性集。

定义5 加权模糊关联规则 $X\Rightarrow Y$ 的加权模糊支持率定义为 $WFSup=\sum_{i=1}^{p+q}w_{y_i}\times(\sum_{j=1}^{n}\prod_{i=1}^{p+q}t_j(y_i))/n$。

定义6 加权模糊关联规则"$X\Rightarrow Y$"的加权模糊信任度定义为 $WFconf=(\sum_{j=1}^{n}\prod_{i=1}^{p+q}t_j(y_i))/\sum_{j=1}^{n}\prod_{i=1}^{p}t_j(y_i)$。

定义7 设 $X=\{y_1,\cdots,y_p\}\subset I$，如果不考虑模糊属性的权重，定义 X 的模糊支持率为 $FSup(X)=(\sum_{j=1}^{n}\prod_{m=1}^{p}t_j(y_m))/n$。

加权模糊支持率的定义由两部分的乘积组成，一部分是规则中所含模糊属性的权值之和 $\sum_{i=1}^{p+q}w_{y_i}$，另一部分是 $(\sum_{j=1}^{n}\prod_{i=1}^{p+q}t_j(y_i))/n$。权值之和 $\sum_{i=1}^{p+q}w_{y_i}$ 能考虑模糊属性的权重和规则中含有模糊属性的数量。规则中含有的模糊属性数量越多，则 $(\sum_{j=1}^{n}\prod_{i=1}^{p+q}t_j(y_i))/n$ 越小，加权模糊支持率就越小；另一方面，规则中含有的模糊属性数量越多，则 $\sum_{i=1}^{p+q}w_{y_i}$ 越大，加权模糊支持率就越大。因此加权模糊支持率能很好地考虑这两方面的平衡。

定理1 如果 $FSup(X)$ 不小于加权模糊最小支持率，则对 X 的任意非空子集，不妨设为 $Y=\{y_1,\cdots,y_l\}$，$l<p$，有 $FSup(Y)=(\sum_{j=1}^{n}\prod_{m=1}^{l}t_j(y_m))/n$ 也不小于加权模糊最小支持率。

从定义4中很容易得到：如果一个模糊属性集是加权模糊大属性集，那么它的子集不一定是加权模糊大属性集。因此，不能直接应用 Apriori 算法来发现加权模糊大属性集。在权重归一化的情况下，由于模糊属性集的加权模糊支持率总是小于模糊支持率，于是就得到第二种加权模糊关联规则的挖掘算法。

第二种加权模糊关联规则的挖掘算法：

输入：数据库 $T=\{t_1,\cdots,t_n\}$，加权模糊最小支持率，加权模糊最小信任度。

输出：加权模糊关联规则

1 应用 FCM 算法将数量型属性离散化，并将记录在属性上的取值划分成若干个模糊集等级；

2 通过数据库 $T=\{t_1,\cdots,t_n\}$ 构造一个新的数据库，新数据库以属性不同的模糊集等级作为数据库的模糊属性；

3 在构造的新数据库上，首先不考虑模糊属性的权重，由于定理1成立，因此可以类似采用 Apriori 算法中找大属性集的方法，找出模糊支持率不小于用户给定加权模糊最小支持率的属性集。由于加权模糊支持率总是小于模糊支持率，这些属性集组成的集合是加权模糊大属性集组成集合的超集；

4 计算这个超集中所有属性集的加权模糊支持率，并把加权模糊支持率小于加权模糊最小支持率的属性集删除，从而得到所有的加权模糊大属性集。第二步的计算中无需再扫描数据库；

5 利用加权模糊大属性集生成所需的加权模糊关联规则，生成算法与 Apriori 算法类似。

结束语 已有数量型关联规则的挖掘算法平等一致地处理各个数量型属性，然而，数据库中的不同数量型属性往往具有不同的重要性。这类数据库包括两种类型，因此文中相应地提出两种加权模糊关联规则的挖掘算法。第一种挖掘算法能有效地考虑数量型属性的权重，并且认为规则的重要性不随着规则中所含属性数量的增加而增加。第二种挖掘算法不仅需要考虑属性的权重，并且认为规则的重要性随着规则中所含属性数量的增加而增加。

参考文献

1 Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proc. of the 1994 International Conf. on Very Large Databases. Santiago, Chile, 1994. 487～499

2 Srikant R, Agrawal R. Mining quantitative association rules in large relational tables. In: Proc. of the ACM-SIGMOD Conf. on Management of Data. Montreal, Canada, 1996. 1～12

3 Chan M K, Ada F, Man H W. Mining fuzzy association rules in database. In: Proc. of the ACM Sixth International Conf. on Information and Knowledge Management. Las Vegas, Neveda, 1997. 10～14

4 陆建江，宋自林，钱祖平. 挖掘语言值关联规则. 软件学报，2001，12(4)：607～611

5 陆建江，钱祖平，宋自林. 正态云关联规则在预测中的应用. 计算机研究与发展，2000，37(11)：1317～1320

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