# 一种语言值关联规则刷洗方法*)

江建华[1] 陆建江[1,2]

(解放军理工大学通信工程学院 南京210007)[1] (东南大学计算机科学与工程系 南京210096)[2]

**摘 要** 通过给定的最小支持率和最小信任度来挖掘语言值关联规则往往会得到很多规则，而且这些规则之间存在一定的冗余，因此需要对语言值关联规则进行筛选。提出一种基于遗传算法的语言值关联规则筛选方法。此方法首先对语言值关联规则进行二进制编码，并通过遗传算法全局搜索一组语言值关联规则，使得对所有样本点的线性平均误差最小。实验表明，算法能够大量减少语言值关联规则的数量，并筛选出对用户更有用的语言值关联规则

**关键词** 数据挖掘，语言值，关联规则，遗传算法，模糊聚类

## A Method for Selecting Association Rules with Linguistic Terms

JIANG Jian-Hua[1] LU Jian-Jiang[1,2]

(Institute of Communications Engineering, PLA University of Science and Technology, Nanjing 210007)[1]

(Department of Computer Science and Engineering, Southeast University, Nanjing 210096)[2]

**Abstract** Using the given minimum support and minimum confidence, a large set of association rules with linguistic terms can be discovered. Furthermore, these rules are usually redundant, so the association rules with linguistic terms should be filtered. An approach of genetic algorithm for selecting association rules with linguistic terms is presented. In this approach, the association rules with linguistic terms are coded by binary code, and genetic algorithm is applied to search a group of association rules with linguistic terms wholly, which make mean linear error least on all samples. The experiment shows that the approach can not only reduce the number of association rules with linguistic terms, but also select the more useful association rules with linguistic terms to the users.

**Keywords** Data mining, Linguistic terms, Association rules, Genetic algorithm, Fuzzy clustering

## 1 引言

Agrawal R 等提出布尔型关联规则的 Apriori 算法[1]。数量关联用来描述数量型属性之间的相互关系，Srikant R 等提出基于支持度的部分 K 度完全的离散化方法[2]，挖掘算法将数量型属性划分成多个区间，区间划分的方法存在划分边界过硬的缺点，同时，在处理高偏度数据时，区间划分的方法很难有效地体现数据的实际分布情况。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]用模糊聚类算法将数量型属性划分成若干个语言值，并提出语言值关联规则的挖掘算法。文[5]改进文[4]的工作，使得挖掘算法能适用于较多属性的数据库。文[5]挖掘语言值关联规则往往会得到很多规则，而且这些规则之间存在一定的冗余，因此需要对规则进行筛选。遗传算法是一种全局优化搜索方法，本文应用遗传算法来筛选规则。

## 2 语言值关联规则的挖掘算法

设 $T=\{t_1,t_2,\cdots,t_n\}$是一个数据库，$t_j$ 表示 $T$ 的第 $j$ 个记录，$I=\{i_1,i_2,\cdots,i_m\}$表示属性集，$t_j[i_k]$表示第 $j$ 个记录在属性 $i_k$ 上的取值。挖掘语值关联规则需要将记录在属性上的取值划分成若干个语言值。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语言值，仍记为 $A_1$与 $A_2$，语言值 $A_1$的隶属度定义为：当取值为 $A_1$时为1，取值为 $A_2$时为0，采用同样的方法可以定义语言值 $A_2$。类别型属性可以采用类似的方法划分。如果是数量型属性，则采用模糊 $c$-均值(FCM)算法[5]划分。语言值关联规则的挖掘算法需通过构造一个新数据库，属性的语言值取为新数据库的属性，记录在属性上的取值按下面方法得到：不妨看语言值属性 $i_1(1)$，第 $j$ 个记录在属性 $i_1(1)$上的取值为原数据库中第 $j$ 个记录在属性 $i_1$上的取值在 $i_1(1)$上的隶属度。记所有语言值属性组成的集合为 $I$，所有记录组成的集合为 $T$，容易知道记录在属性上的取值都属于区间[0,1]。

在新数据库中，设 $X=\{y_1,y_2,\cdots,y_p\}$，$Y=\{y_{p+1},y_{p+2},\cdots,y_{p+q}\}$是 $I$ 的子集，$X\cap Y=\phi$，讨论的语言值关联规则为"$X\Rightarrow Y$"。定义语言值属性集 $X$ 的支持率为 $Sup(X)=(\sum_{j=1}^{n}\prod_{m=1}^{p}t_j[y_m])/n$。如果 $Sup$

*)基金项目：国家自然科学基金青年科学基金(60303024)。江建华 硕士生，主要研究领域为数据挖掘。

$(X)$不小于用户给定的最小支持率，则称 $X$ 为频繁语言值属性集。规则的支持率定义为 $Sup=(\sum_{j=1}^{n}\prod_{m=1}^{p+q}t_j(y_m))/n$，信任度定义为 $Conf=Supsup(X)$。有意义的语言值关联规则可以从频繁语言值属性集中直接生成。由语言值属性集 $X$ 的支持率定义容易知道，如果 $X$ 是频繁语言值属性集，那么它的所有非空子集都是频繁语言值属性集。由于这个性质的保证，就很容易改进 Apriori 算法[1]发现所有的频繁语言值属性集。进一步的细节可见文[5]。

## 3 语言值关联规则的刷洗方法

设挖掘得到 $l$ 条语言值关联规则 $R_j$。如果 $i_1$ is $f_1^j$，且 $i_2$ is $f_2^j$，…，且 $i_{m-1}$ is $f_{m-1}^j$，则 $i_m$ is $f_m^j$，$j=1,2,\cdots,l$。其中 $f_k^j$ 是语言值。应用 $l$ 条规则可以进行预测，预测采用文[6]中等权重的标准可加性模型。当前件的输入值为 $x_1,x_2,\cdots,x_{m-1}$ 时，得到后件 $i_m$ 的预测结果为：

$$y_o=\frac{\sum_{j=1}^{l}a_jv_jc_j}{\sum_{j=1}^{l}V_ja_j}$$

其中 $a_j=\prod_{k=1}^{m-1}f_k^j(x_k)$ 表示规则的激活程度，$c_j=\frac{\int_R xf_m^j(x)\mathrm{d}x}{\int_R f_m^j(x)\mathrm{d}x}$ 是规则后件输出语言值的形心，$V_j=\int_R f_m^j(x)\mathrm{d}x$，$l$ 是规则的数目。

语言值关联规则的刷洗准则是：从 $l$ 条语言值关联规则中刷洗出一定数量的规则来改变输出值 $y_{oi}$，使得线性平均误差 $E=\frac{1}{n}\sum_{i=1}^{n}|y_{oi}-y_i|$ 最小，其中 $y_i$ 是实际值，$n$ 是本总数。

遗传算法[7]筛选规则是一个组合优化问题，下面是遗传算法执行过程。

(1)*编码* 采用二进制编码，通过编码生成固定长度的染色体。方法如下：考虑具有 $l$ 条规则 $R_j$，$j=1,2,\cdots,l$ 的初始规则库，设刷洗过程中的规则库为 $B^F$，那么可以用一个 $l$ 位的二进制位串 $S=(s_1,s_2,\cdots,s_l)$ 来表示规则库 $B^F$，它满足如果 $s_j$ 为1，则 $R_j\in B^F$，否则 $R_j\notin B^F$。

(2)*初始群体的生成* 初始群体的第一个个体通过引入一个表示完整的初始规则集的染色体生成，也就是每个基因 $s_j$ 都等于1的染色体。其余的个体通过随机方法产生，即每个个体的基因随机取为0或1。

(3)*适应度的设定* 设 $S$ 为初始规则集的一个子集，其所含规则数为 $N(S)$，预测的绝对误差和为 $N(S)$。筛选的目标为最小化 $E(S)$ 和 $N(S)$，这是一个两目标的组合优化问题。引入权值 $0<\omega<1$，定义规则集 $S$ 的适应度 $f(S)$ 为：

$$f(S)=\begin{cases}1/(\omega\cdot\frac{E(S)}{E_0}+(1-\omega)\cdot\frac{N(S)}{N_0}) & \text{if } E_0\neq 0\\ 1/(\omega\cdot E(S)+(1-\omega)\cdot\frac{N(S)}{N_0}) & \text{if } E_0=0\end{cases}$$

其中 $E_0$ 是使用初始规则库预测的绝对误差和，$N_0$ 是初始规则库包含的规则数目。对于当前代群体中的个体，采用上述适应度函数进行评估。

(4)遗传操作

1)选择操作。个体选择采用随机遍历采样过程并结合最佳个体保存的选择机制。首先依据各个体的适应度大小进行排序。设群体大小为 $N$，适应度最大的序号为1，适应度最小的序号为 $N$；然后给每个个体赋以一个与其序号成线性关系的新的适应度值，序号最小的个体所取的新适应度值设为 $f_{max}$，取值范围为[1,2]，序号最大的个体取适应度值为 $f_{min}=2-f_{max}$，序号在中间的个体其适应度按照公式 $f(i)=f_{max}-(i-1)\cdot(f_{max}-f_{min})/(N-1)$ 计算。随机遍历采样过程可以看作是赌轮方法的一个变形。个体在赌轮中对应的区域跨过某指针位置时，则该个体被选上。最佳个体保留选择机制把当前群体中适应度最高的个体完整地复制到下一代群体中。

2)交叉操作。个体基因重组使用标准的二进制两点交叉操作，生成下一代个体群。

3)变异操作。对生成的各个体，按变异概率进行基本变异操作。变异的基因由0变为1或由1变为0。

## 4 实验分析

实例数据库 Abalone 取自 UCI，数据库有4177条记录，8个数量型属性，记为 $i_1,i_2,\cdots,i_8$。文中讨论前件为 $i_1,i_2,\cdots,i_7$，后件为 $i_8$ 的规则，实验中采用 FCM 算法将数量型属性划分为3个三角模糊数：大、中、小，分别记为 $i_m(1),i_m(2),i_m(3)$，$m=1,2,\cdots,8$。表1中给出了部分三角模糊数的参数，挖掘语言值关联规则的形式为：

如果 $i_1$ is $i_1(l_1)$ 且 $i_2$ is $i_2(l_2)$，…，且 $i_7$ is $i_7(l_7)$ 则 $l_8$ is $i_8(l_8)$。$l_m=1,2,3$，$m=1,2,\cdots,8$

为了全面了解前7个属性对年龄的影响，需要挖掘当后件取为3个不同语言值的规则。但由于 Abalone 数据库中数据分布不均匀，当3个不同语言值的规则全部出现时，得到的语言值关联规则往往很多，而且这些语言值关联规则之间存在一定的冗余，因此需要对语言值关联规则进行筛选。取最小支持率为0.000226和最小信任度为0.5，挖掘得到300条语言值关联规则，用300条语言值关联规则进行预测，所有样本点上的平均线性误差为3.679。按照上面的方法采用遗传算法筛选语言值关联规则，筛选时遗传参数的最大代数设为500，群体大小为61，交

(下转第50页)

4)匹配器 匹配器的主要任务有两点:一是将服务发布者提交过来的服务语义文件和本体文件保存在本地同时在语义文件数据库中追加相应的记录指向这两个文件的保存地点;二是接受查询重写器提交过来的查询语句,然后将其作为参数调用语义匹配程序在语义数据库中进行语义匹配查询,构造查询结果并将其返回给 UDDI 扩展接口层,UDDI 扩展接口层将结果再返回给服务的请求者。

在进行查询匹配的时候,这里主要实现的是输入输出参数的匹配,即比较请求的输入输出参数和发布服务的输入输出参数所属的本体类之间的关系,符合一定的关系就认为是不同程度的匹配。

## 4 系统特点

近年来,Web 服务技术已成为互联网和分布式计算领域的一个新的研究热点,而其中的 Web 服务的智能发现和选择是其中的核心技术之一。本文提出的基于 Agent 和 DAML-S 的 Web 服务优化选择框架正是在这个领域的一个探索。框架经过原型化的实现已经取得了一定的成果,主要表现在以下几个方面:1)借鉴了人类社会的决策方法,加入了 Agent 推荐服务机制,使得服务选择的最终结果接近于最优。2)设计并实现了基于 DAML-S 语义匹配查询算法,能够较好地实现服务请求和发布服务之间在输入输出接口上基于语义的匹配。

**结束语** 本文设计开发了一个基于 Agent 和 DAML-S 的 Web 服务选择框架。将该框架应用于一个网上汽车销售系统,证明本文所提出的框架有一定的商业价值。由于 Web 服务的选择才刚刚起步,还有很多不足,比如服务质量的可扩展和跨领域[5]的问题,我们可以考虑建立一个服务质量标准或模型来完善等等,因此 Web 服务的选择问题有很好的发展前景。

## 参 考 文 献


1 Barber S K, Kim J. Belief revision process based on trust: Agents evaluating reputation of information sources. In: R. Falcone, M. P. Singh, Y.-H. Tan, eds. Trust in Cybersocieties, volume 2246 of LNAI, Springer-Verlag, 2001. 73～82

2 Banerjee A. C# Web Services. 清华大学出版社, 2002

3 DAML Services Coalition. DAML-S versions 0.5, 0.6 and 0.7. http://www.daml.org/services/

4 林弘之. Web Services 原理与开发实务. 电子工业出版社, 2003

5 Liu Yutu. QoS Computation and Policing in Dynamic Web Service Selection. www.cs.swt.edu/～hn12/papers/ p494-ngu/ p494-ngu.html. 2004


(上接第28页)

又概率为0.6,个体变异概率为0.1,ω 为0.8。筛选后得到84条语言值关联规则,在所有样本点上的平均线性误差为2.336。结果是比原来减少了216语言值关联规则,但误差却减少1.343。

表1 部分三角模糊数的参数

| 大 | 中 | 小 | |
|---|---|---|---|
| 长度 | (0.500,0.633,1.198) | (0.335,0.501,0.632) | (－0.374,0.330,0.490) |
| 直径 | (0.387,0.498,0.969) | (0.253,0.387,0.499) | (－0.313,0.247,0.382) |
| 高度 | (0.130,0.182,1.674) | (0.089,0.139,0.181) | (－0.139,0.088,0.143) |

**结论** 提出一种基于遗传算法的语言值关联规则筛选方法。此方法首先对语言值关联规则进行二进制编码,并通过遗传算法全局搜索一组语言值关联规则,使得对所有样本点的线性平均误差最小。实验表明,算法能够大量减少语言值关联规则的数量,并筛选出对用户更有用的语言值关联规则。今后的努力方向是使得平均线性误差进一步缩小,初步的设想是调整语言值关联规则中三角模糊数的参数。

## 参 考 文 献


1 Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules. In: Proc. of the 1994 Intl. Conf. on Very Large Databases. Santiago de Chile, Chile: Morgan Kaufmann Press, 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, Quebec, Canada: ACM Press, 1996. 1～12

3 Chan M K, Ada F, Man H W. Mining Fuzzy Association Rules in Database. SIGMOD Record, 1998, 27(1): 41～46

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 邹晓峰,陆建江,宋自林. 语言值关联规则挖掘算法. 系统仿真学报, 2002, 14(9): 1130～1132

6 黄崇福译. 模糊工程. 西安:西安交通大学出版社, 1999. 36～39

7 Goldberg D. Genetic Algorithms in Search, Optimization and Machine Learning. New York: Addison-Wesley, 1989