文章编号: 1001-0920(2008)08-0957-04

基于免疫原理的模糊关联规则挖掘算法

张 雷[1,2], 李人厚[1]
(1. 西安交通大学 系统工程研究所, 西安 710049; 2. 河南科技大学 电子与信息工程学院, 河南 洛阳 471003)

摘 要: 提出一种基于免疫原理的人工免疫算法,用于模糊关联规则的挖掘. 该算法通过借鉴生物免疫系统中的克隆选择原理来实施优化操作,它直接从给出的数据中,通过优化机制自动确定每个属性对应的模糊集合,使推导出的满足条件的模糊关联规则数目最多. 将实际数据集和相关算法进行性能比较,实验结果表明了所提出算法的有效性.

关键词: 关联规则; 数据挖掘; 模糊集合; 免疫原理
中图分类号: TP18 **文献标识码**: A

Algorithm for mining fuzzy association rules based on immune principles

ZHANG Lei[1,2], *LI Ren-hou*[1]
(1. Institute of System Engineering, Xi'an Jiaotong University, Xi'an 710049, China; 2. School of Electronics and Information Engineering, He'nan University of Science and Technology, Luoyang 471003, China. Correspondent: ZHANG Lei, E-mail: leizhang87@163.com)

Abstract: An algorithm is proposed for mining fuzzy association rules based on immune principles, which is mainly inspired by the clonal selection principle of biological immune systems. It is employed to optimize the number of strong rules that satisfy the specified thresholds by adjusting the parameters of fuzzy sets for each quantitative attribute. The performances of the algorithm is compared with other relevant algorithms and the experimental results show the effectiveness of the algorithm.

Key words: Association rules; Data mining; Fuzzy sets; Immune principles

1 引 言

关联规则挖掘是数据挖掘技术中一个重要的研究方向. Agrawal 等[1]首先提出了用于购物篮分析的布尔型关联规则挖掘问题. 关联规则挖掘的任务是发现事务型数据库中属性或项之间相互关联的性质. 关联规则挖掘的目标是寻找所有支持度和确信度高于用户所设定阈值的关联规则.

布尔型关联规则关注的是一项是否包含在一个事务中,而没有考虑其数量信息. 但是在大多数应用中,数据库中都包含有数量型属性,这时可通过转化将其变为布尔型关联规则的挖掘问题,所得到的关联规则称为数量型关联规则[2]. 解决的方法是将数量型属性划分为离散的区间,并用新的布尔型属性替代原属性. 但这种方法存在边界划分过硬的问题,即对于区间边界的元素或者忽视或者过分强调.

为解决边界划分过硬问题,近年来,人们将模糊集合的概念引入关联规则中,并提出了相应的模糊关联规则挖掘算法[3-5]. 文献[3]应用聚类算法 CLARANS[6]确定模糊关联规则中每个属性所对应的模糊划分,它根据聚类中心确定模糊集合的参数. 文献[4]应用一种更为有效的 CURE 聚类算法,来确定每个属性所对应的模糊集合. 文献[5]则引入了关联规则的模糊有趣性和确信度的概念.

针对模糊关联规则挖掘,在已有的大多数方法中,属性所对应的模糊集合或者由专家给出,或者通过聚类方法进行确定. 利用专家提供模糊集合只适用于一定的场合,因为不可能总得到相关的专家. 而应用聚类方法确定模糊集合,结果往往也不理想,因为它是基于数据的分布特征来确定模糊集合,与关联规则中的相关概念没有必然的联系.

收稿日期: 2007-05-20; **修回日期**: 2007-11-07.
基金项目: 国家科技攻关项目(2005BA115A01); 国家自然科学基金项目(60373135).
作者简介: 张雷(1974—),男,河南洛阳人,博士生,从事智能计算、数据挖掘等研究; 李人厚(1935—),男,浙江宁波人,教授,博士生导师,从事 CSCW,智能控制等研究.

在模糊关联规则的挖掘过程中,确定模糊集合的参数可视为一种优化问题,优化的目标可以是所得到关联规则的数目和有趣性等. 文献[7]提出了一种数量型关联规则的挖掘算法,它通过优化来确定属性对应的区间,使得关联规则的支持度或确信度取得最优值. 人工免疫系统是一种新型的计算智能范例,它通过模拟自然免疫原理和功能解决实际问题,在优化、数据分析和机器人技术等许多领域得到了广泛应用[8]. 本文提出一种基于免疫原理的人工免疫算法,它能利用给出的数据自动确定属性对应的模糊集合,使得满足阈值条件的关联规则的数目最多. 将实际数据集和相关算法进行比较,实验结果表明,本文算法具有更优的性能.

2 模糊关联规则

假设 T 是一个包含 n 个事务的数量型数据库,其所有属性的集合为 I. 本文算法所采用的模糊关联规则的类型为

$$\begin{aligned}&\text{if } X=\{x_1,x_2,\cdots,x_p\} \text{ is } A=\{f_1,f_2,\cdots,f_p\},\\&\text{then } Y=\{y_1,y_2,\cdots,y_q\} \text{ is } B=\{g_1,g_2,\cdots,g_q\}.\end{aligned}\tag{1}$$

其中:X 和 Y 为互不相交的属性集合,A 和 B 分别为 X 和 Y 中包含属性所对应的模糊集合,"X is A"称为关联规则的前件,而"Y is B"则称为规则的后件.

定义 1 项目(简称项) (a_i,f_i). 其中:a_i 为属性集合 I 中的某个属性,f_i 为属性 a_i 对应的模糊集合. 称 $(D,F)=(a_1,f_1)\wedge(a_2,f_2)\wedge\cdots\wedge(a_k,f_k)$ 为长度为 k 的项目集合(简称项目集). 其中:D 为属性的集合,F 为 D 中所有属性对应的模糊集合.

定义 2 模糊支持度. 项目集合 (D,F) 的模糊支持度定义为

$$\sup(D,F)=\frac{\sum_{t_i\in T}\prod_{d_j\in D}\mu_{f_j}(t_j[d_j])}{|T|}.\tag{2}$$

其中:f_j 为属性 d_j 对应的模糊集合,μ_{f_j} 为 f_j 所对应的隶属度函数,$t_i[d_j]$ 为数据库中第 i 个数据在属性 d_j 上的值,$|T|$ 为数据库中的数据数目.

当项目集合的模糊支持度不小于用户设定的阈值时,便称其为频繁项目集合.

由定义 2,式(1)所表示的关联规则 R 的支持度和确信度的定义分别为

$$\mathrm{FS(R)}=\sup(Z,C),\tag{3}$$

$$\mathrm{FC(R)}=\frac{\sup(Z,C)}{\sup(X,A)}.\tag{4}$$

其中:$Z=X\cup Y$, $C=A\cup B$.

模糊关联规则的挖掘过程可描述为:首先,基于属性对应的模糊集合和用户设定的阈值,找到所有长度大于或等于 2 的频繁项目集;然后对于每个频繁项目集,产生所有可能的模糊关联规则,并从中选择支持度和确信度都满足阈值条件的规则.

当关联规则的支持度和确信度都满足设定阈值时,称其为强关联规则. 某些强关联规则可能包含属性的负相关,这时所得到的规则是无意义的. 本文引入一种规则的有趣性度量,作为一种过滤标准,以删除这些负相关的规则. 形如式(1)的关联规则 R 的有趣性定义为[6]

$$I(\mathrm{R})=\frac{\sup(Z,C)}{\sup(X,A)\times\sup(Y,B)}.\tag{5}$$

当规则的有趣性的值小于 1 时,该规则便包含属性的负相关.

3 基于免疫原理的模糊关联规则挖掘算法

挖掘模糊关联规则,首先要确定每个属性所对应的模糊集合. 选取的模糊集合是否合适,对于所得到的模糊关联规则的质量至关重要. 模糊集合可由专家确定,或由给出的数据集自动提取. 本文所提出的算法能由给出的数据集自动优化确定属性所对应的模糊集合,使得满足阈值条件的关联规则的数目最多.

本文算法主要借鉴了免疫系统中的克隆选择和超变异原理,克隆选择属于自然选择的一种类型,抗原对抗体种群具有选择作用,那些能够有效识别抗原的抗体将被选择进行克隆增殖和超变异.

3.1 个体的编码

算法中每个抗体表示所有属性对应的隶属度函数的参数. 这里采用三角形隶属度函数,因为这是一种最为常用且简单的形式.

假定每个属性包含 5 个模糊集合,它们可以用 3 个参数进行表示. 例如,对于属性 i,其对应的的隶属度函数如图 1 所示. 当属性的数目为 n 时,一个抗体的编码可表示为如下形式:

$$p_1^1\ p_2^1\ p_3^1\ p_1^2\ p_2^2\ p_3^2\ \cdots p_1^n\ p_2^n\ p_3^n.$$

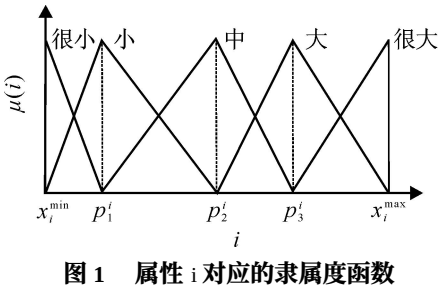

图 1 属性 i 对应的隶属度函数

3.2 抗体的适应值

适应值用于评价抗体的优劣. 选取恰当的适应度函数是本文算法的一个关键问题,因为算法的主要目标就是优化种群中抗体的适应值. 针对每个抗体所给出的模糊集合,确定所有满足阈值条件的关

联规则,并删除负相关的规则,所得到的关联规则的数目定义为该抗体的适应值,即

$$\mathrm{fit}(i) = \sum_{j=1}^{N_R} B(\mathrm{R}_j). \tag{6}$$

其中:$\mathrm{fit}(i)$ 为抗体 i 的适应值;N_R 为所得到的关联规则的数目;$B(\mathrm{R}_j)$ 为标志位,用于判断规则 R_j 是否满足设定的阈值条件,其定义为

$$B(\mathrm{R}_j) = \begin{cases} 1, & \mathrm{FS}(\mathrm{R}_j) \geq \mathrm{minsup}, \\ & \mathrm{FC}(\mathrm{R}_j) \geq \mathrm{minconf} \text{ 且 } I(\mathrm{R}_j) \geq 1; \\ 0, & \text{其他}. \end{cases} \tag{7}$$

这里 minsup 和 minconf 分别表示最小支持度和最小确信度.

本文算法具体步骤如下:

1) 确定种群规模 N,并生成初始抗体种群.

2) 计算种群中每个抗体的适应值.

3) 从当前种群中选择适应值最高、且适应值互不相同的 N_1 个抗体,进行克隆操作. 每个抗体克隆的数目 N_c 与其适应值成正比.

4) 对于克隆新生成的抗体,实施超变异操作. 抗体的适应值越高,其对应的变异率越小.

5) 从种群中未进行克隆的个体中随机选择 N_d 个抗体,用克隆新生成的抗体中适应值最高的 N_d 个抗体进行替换,得到下一代抗体种群.

6) 若迭代周期达到指定迭代次数,则算法结束;否则转步骤 2).

该算法通过随机的方式生成初始抗体种群. 假定所有属性都采用 5 个隶属度函数,并可用 3 个参数来确定,如图 1 所示. 对应每个抗体,从每个属性的变化范围内随机选择 3 个实数,作为该属性对应的隶属度函数参数,即

$$x_j^{\min} < p_1^j < p_2^j < p_3^j < x_j^{\max},\ j = 1,2,\cdots,n.$$

本文从种群中选择适应值最高、且适应值互不相同的 N_1 个抗体,作为进行克隆增殖的候选个体. 这种选择策略能够保证搜索区域的多样性,避免陷入未成熟收敛. 每个抗体克隆的数目 N_c 与其适应值成正比,即

$$N_c = \max\{N_{\min}, \mathrm{round}(N_{\max} \times \mathrm{fit}(i)/f_{\max})\}. \tag{8}$$

其中:$N_{\min}$ 和 $N_{\max}$ 分别为最小和最大克隆数目,$\mathrm{fit}(i)$ 为抗体 i 的适应值,$f_{\max}$ 为当前种群中抗体的最大适应值.

每个克隆新生成的个体还要进行一种超变异的过程. 抗体的适应值越高,其变异率则越低,即

$$\mathrm{mr}(i) = (1 - \mathrm{fit}(i)/f_{\max})(\mathrm{mr}_{\max} - \mathrm{mr}_{\min}) + \mathrm{mr}_{\min}, \tag{9}$$

其中 $\mathrm{mr}_{\max}$ 和 $\mathrm{mr}_{\min}$ 分别为最大和最小变异率.

由于采用实数编码方法,对于每个进行变异的基因,可按下式进行变异操作:

$$g(i,j) = g(i,j) + \gamma_i\rho,$$
$$x^{\min} \leq g(i,j) \leq x_i^{\max}. \tag{10}$$

其中:$g(i,j)$ 为第 i 个属性的第 j 个参数,γ_i 为第 i 个属性的最大变异量,$\rho \in [-1,1]$ 为一个均匀分布的随机数.

在更新抗体种群时,随机选择 N_d 个未进行克隆操作的抗体,用新生成的抗体中适应值最高的 N_d 个抗体进行替换. 每代种群的规模保持恒定. 算法在选择克隆候选抗体时,既考虑了抗体的适应值,又考虑到抗体之间的相似性,该过程包含了对相似抗体的抑制,它能与种群更新机制共同保持种群中个体的多样性.

4　仿真实验

针对 UCI 机器学习数据库[9] 中的 Wisconsin Breast cancer 数据集进行仿真实验,以检验所提出算法的性能,并与文献[4] 中基于聚类的方法进行性能比较. 实验中,每个属性对应的模糊集合的数目设为 5 个,种群包含 50 个抗体,而最大迭代数目设为 200 代.

当最小确信度设为 0.7,最小支持度从 0.5 增加到 0.6 时,两种方法所得到的满足阈值条件的关联规则数目如图 2 所示. 可以看出,随着最小支持度的增加,频繁项集的数目逐渐减少,所得到的关联规则数目也会相应减少. 但对于不同的阈值条件,本文算法都能得到更多的满足阈值条件的关联规则.

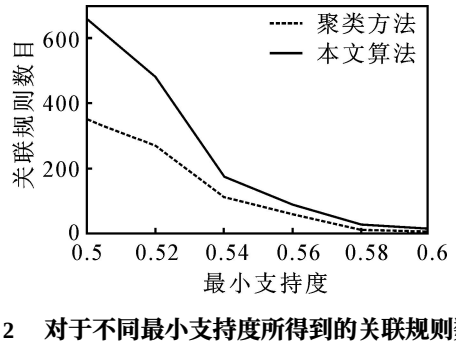

图 2　对于不同最小支持度所得到的关联规则数目

当最小支持度设为 0.5,最小确信度从 0.6 增加到 0.85 时,两种方法所得到的满足阈值条件的关联规则数目如图 3 所示. 可以看出,本文算法仍能得到更多的满足条件的关联规则.

最后对算法的收敛性进行实验,即测试所得到的满足阈值条件的关联规则的数目随迭代周期的变化情况. 当最小支持度取 0.54,最小确信度取 0.7 时,实验结果如图 4 所示. 可以看到,算法在演化到 200 代左右时开始收敛.

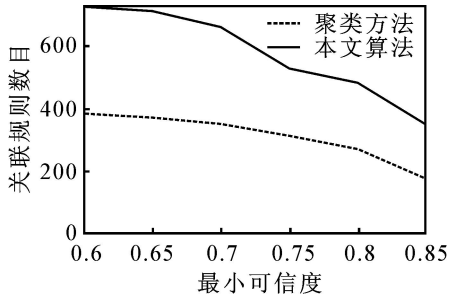

图 3 对于不同最小确信度所得到的关联规则数目

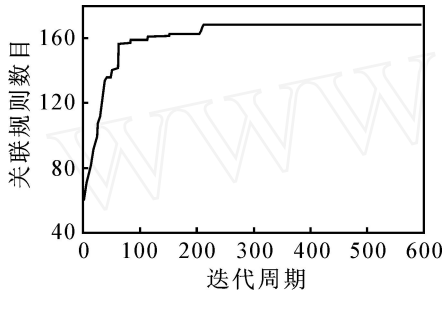

图 4 所得到的关联规则数目与迭代周期的关系

5 结 论

本文提出一种基于免疫原理的人工免疫算法,用于模糊关联规则的挖掘. 在关联规则的挖掘过程中,虽然可以利用专家来确定每个属性对应的模糊集合参数,但这只适用于一定的场合. 本文算法则直接从给出的数据中,通过借鉴免疫原理自动确定每个属性对应的模糊集合. 将实际数据集与相关算法进行了比较,实验结果表明了本文算法具有更优的性能.

参考文献(References)

[1] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[C]. Proc of ACM SIGMOD. Paris, 1993: 207-216.

[2] Srikant R, Agrawal R. Mining quantitative association rules in large relational tables [C]. Proc of ACM SIGMOD. Montreal, 1996: 1-12.

[3] Fu A W C, Wong M H, Sze S C, et al. Finding fuzzy sets for the mining of association rules for numerical attributes[C]. Proc of the Int Symposium of Intelligent Data Engineering and Learning. Hong Kong, 1998: 263-268.

[4] Kaya M, Alhajj R, Polat F, et al. Efficient automated mining of fuzzy association rules[C]. Proc of the Int Conf on Database and Expert Systems with Applications. Aix-en-Provence, 2002: 133-142.

[5] Krishna K S, Krishna P R, De S K. Discovering fuzzy association rules with interest and conviction measures [C]. Knowledge-based Intelligent Information and Engineering Systems. Melbourne, 2005: 101-107.

[6] Ng R, Han J. Efficient and effective clustering methods for spatial data mining[C]. Proc of the Int Conf on Very Large Databases. Santiago, 1994: 144-155.

[7] Fukuda T, Morimoto Y, Morishita S, et al. Data mining using two dimensional optimized association rules: Scheme, algorithms and visualization[C]. Proc of ACM SIGMOD. Montreal, 1996: 13-24.

[8] de Castro L N, Timmis J. An introduction to artificial immune systems: A new computational intelligence paradigm[M]. New York: Springer-Verlag, 2002.

[9] Blake C, Merz C. UCI repository of machine learning databases[EB/OL]. http://www.ics.uci.edu/mlearn/MLRepository.html, 1998.

2008 中国控制与决策会议胜利召开

本刊讯 2008 中国控制与决策会议(2008 CCDC),已于 7 月 2 日～4 日在山东省烟台市召开. 会议由东北大学和 IEEE 新加坡工业电子分会联合主办,鲁东大学具体承办. 来自国内外高等院校和科研机构的 500 多位代表参加了会议,其中国外和海外的代表 30 余人. 这是一次国际学术盛会,大家齐聚一堂,交流学术思想,讨论学术问题,充满了浓厚的学术气氛.

本届会议邀请了 5 位国内外著名专家学者和 7 位国内外著名教授,就当前控制与决策领域的热点问题和最新研究成果作了专题大会报告和准大会报告. 这些报告对当前前沿学科的一些热点问题进行了阐述和评论,受到代表们的普遍欢迎.

大会发行了《2008 中国控制与决策会议论文集》光盘,光盘中的 1090 篇论文将由 ISTP 收录,并将进入 IEEE Xplore Data Base,同时将被 EI 检索.