广西科学院学报
Journal of Guangxi Academy of Sciences

2007,23(3):144~146
Vol. 23,No. 3 August 2007

一个多层次模糊规则的逐维挖掘算法*

Dimension Mining Algorithm of Multidimensional and Multilevel Fuzzy Rule

李海滨
LI Hai-bin

(广西民族大学数学与计算机学院,广西南宁 530006)
(School of Mathematics & Computer Science, Guangxi University for Nationalities, Nanning, Guangxi, 530006, China)

摘要:在模糊聚类的基础上,分析多维规则之间模糊传递关系的分布特性,提出一个多层次模糊规则的逐维挖掘算法(MMFCRA)。该算法通过逐步精简候选集的递推计算来生成多维模糊规则集,从而建立多维多层次的模糊规则挖掘模型,能够在不减少维数的情况下,既有效地降低了查询计算的次数,又反映了不同维数指标区间组合的影响情况,同时也保证了逐维挖掘算法的收敛性,避免了知识的遗失。

关键词:模糊规则 数据挖掘 金融危机
中图法分类号:O159;TP311 **文献标识码**:A **文章编号**:1002-7378(2007)03-0144-03

Abstract: By analyzing the distribution features of fuzzy transitivity in multidimensional data on the basis of fuzzy clustering, this paper proposes a dimension mining algorithm of multidimensional fuzzy rule (MMFCRA). The algorithm builds a collection of multidimensional fuzzy rules through recursive calculation of reducing candidate intemsets, and subsequently establishes a multidimensional mining model. Without reducing the dimension, this algorithm not only efficiently cuts down the number of inquiry computation, but also reflects the influence of combination of different dimension indexes, and guarantees the convergence of MMFCRA as well, avoiding the loss of knowledge.

Key words: fuzzy rule, data mining, financial crisis

数据挖掘是涉及人工智能和数据库等多学科的一门当前相当活跃的研究领域,并已经应用在许多领域中,因此从大量的数据中智能地、自动地提取出有价值的知识和信息的研究,即数据挖掘,具有十分重要的现实意义[1]。多维多层次模糊知识(模糊预测规则)的挖掘是目前数据挖掘的一个研究热点[2,3]。针对多维数据的知识发现情况,受计算能力的限制,直接采用遍历所有的属性组合是很难实现的,因此,人们通常采用降维的方法来实现其数据的挖掘,但是在属性作用机理不能完全精确描述的情况下,这样处理容易遗失一些关键的信息。

本文在模糊聚类的基础上,通过对多维规则之间模糊传递关系的分布特性进行分析,提出模糊规则的逐维挖掘算法(MMFCRA)。该算法通过逐步精简候选集的递推计算来生成多维模糊规则集,从而建立多维多层次模糊规则挖掘模型。该算法在不减少维数的情况下,既有效地降低了查询计算的次数,又反映了不同维数指标区间组合的影响情况,同时也保证了收敛性,避免了知识的遗失。

1 相关研究工作

1.1 模糊规则的挖掘

在数据挖掘中,按照研究的粒度大小可以将规则分为单层次规则和多层次规则,按照挖掘前后的知识描述模式可以将规则分为确定性规则和模糊性规则[1]。由于客观世界的多样性和复杂性,许多系统很难使用精确和确定的概念来表示,因此,对于连续性数据用模糊规则来表示是比较合适的。对客观世

收稿日期:2006-11-09
修回日期:2007-04-06
作者简介:李海滨(1970-),女,讲师,主要从事人工智能,多智能体,知识工程等研究。
*广西自然科学基金项目(0542048)资助。

界的模糊处理，首先是对数据进行模糊化处理，然后再根据模糊值采用数据挖掘的方法来发现其内在规则，最后通过反模糊化的方法生成用户模式的语言描述[4]。在预测规则中，由于未来世界是未知的，因此单一的描述因提供的信息太少而不能满足决策者的需求。本文采用模糊语言来反映给用户，即规则的左侧是指标的聚类描述，右侧是评价集的多层次模糊描述(即模糊结果集)。

1.2 模糊聚类分析

为了在多维数据集中发现模糊知识，需要对数据进行模糊聚类分析，其中，对连续型数值的聚类分析一般是按区间划分来进行。目前主要的聚类分析方法有：K 均值聚类、山峰聚类法、快速聚类法等[2]。数据库中各个属性的数据通常都是数值型的，对这类数据的处理仅用布尔(boolean)型数据类型是不能反映数据间实际联系特征的。本文采用差异度的模糊聚类方法[4]来处理数值型数据，该方法包括两部分：一是对指标集的评价；二是指标集的聚类分析。对指标集的评价首先确定模糊描述集，然后再将模糊描述值赋予各个元组；指标集的聚类分析主要是对各个属性采用聚类分析和专家评价相结合的方法来确定。

2 多维多层次模糊规则挖掘算法

2.1 算法的基本原理

通过对聚类区间的分析可以发现相同的原因可以产生不同的结果(支持度和可信度不同)，结果集的数目随最小支持度($Sup_{\min}$)和最小可信度($Conf_{\min}$)的提高而减少。同时，随着指标的增加，在相同的最小支持度和最小可信度下，挖掘出的规则数目将减少。因此为了挖掘多维规则，需要根据维数情况来动态地确定最小支持度和最小可信度，这样就产生不同维数的规则之间存在的不完全包含传递性。

2.2 算法的实现步骤

在规则挖掘之前，首先对事务数据库进行数据处理，选择主题，再进行聚类分析，产生 n 维模糊聚类数据库 $\tilde{T}^n$。在进行支持度和可信度计算时，当所关注的结果集(兴趣集)的支持度和可信度均大于相应的最小支持度和最小可信度时，就把其加入到规则集或候选集。算法步骤如下。

步骤1 对$\forall\ k(1\leqslant k\leqslant n)$ 确定 k 维的最小支持度、最小可信度、最小候选支持度($CSup^k_{\min}$)和最小候选可信度($CConf^k_{\min}$)。

步骤2 当 $k=1$ 时，计算模糊规则集和聚类候选集 $\tilde{C}^1$。

步骤3 对于 $k>1$，根据 $k-1$ 维聚类候选集 $\tilde{C}^{k-1}$ 和单维聚类候选集 $\tilde{C}^1$ 产生 k 维聚类候选集 $\tilde{C}^k$。

步骤4 根据 $\tilde{C}^k$ 以及 $Sup^k_{\min}$、$Conf^k_{\min}$、$CSup^k_{\min}$ 和 $CConf^k_{\min}$，遍历数据库 $\tilde{T}^n$，生成 k 维模糊规则集 $\tilde{R}^k$ 和 k 维聚类候选集 $\tilde{C}^k$。

步骤5 若 $k=n$，转步骤6；否则，$k=k+1$，转步骤3。

步骤6 输出模糊规则集 $\tilde{R}=\{\tilde{R}^1,\tilde{R}^2,\cdots,\tilde{R}^n\}$。

步骤7 结束。

2.3 算法分析

保证挖掘算法的收敛性是采用逐维双重筛选方法挖掘多维多层次模糊规则的一个主要问题。在规则挖掘过程中，组合查询计算的次数随维数的增加而呈指数级增长。在此，收敛性主要是指随着维数的增加，通过属性候选集的简约和查询算法的优化，可以有效地降低查询的次数和属性聚类的组合数目，从而保证在用户可以容忍的时间内挖掘出相关的模糊规则。在本文算法中，我们通过生成候选集的方法来保证其收敛性。

最小侯选支持度和最小候选可信度的确定关系到是否能挖掘到所期望的完整知识，这不仅与数据聚类有关，也与用户选择挖掘的主题有关。随着维数的增加，最小侯选支持度和最小候选可信度的值也相应地降低，因此可以将 $k=n$ 时计算值反推作为最小候选支持度和最小候选可信度的初始值。取 $CSup^k_{\min}=\lambda P_{发生}$，其中 $P_{发生}$ 为数据集中某事件发生的概率，λ 是为了弥补样本与总体间统计误差而引入的修正系数。$CConf^k_{\min}$ 取相应的最小值。

最小支持度和最小可信度是影响发现规则数目的主要因素，随着 k 的增加，候选集的减缩，$Sup^k_{\min}$ 和 $Conf^k_{\min}$ 的值也相应减小。若保持 $Sup^k_{\min}$ 和 $Conf^k_{\min}$ 不变，则随着维数的增加，规则的数目将逐步减少。$Sup^k_{\min}$ 和 $Conf^k_{\min}$ 的确定与用户模式有关，用户可以通过逐步改变 $Sup^k_{\min}$ 和 $Conf^k_{\min}$ 的数值，最后得到满意的挖掘模式和模糊规则集。

3 应用实例

在金融危机预警系统中，影响金融危机的指标有很多种，这里采用常用的14个指标[3]进行聚类分析，见表1。

表1 各指标聚类分析结果

指标名称	聚类数目	指标名称	聚类数目
GDP增长率	5	短期外债/外债总额	8
通货膨胀率	9	外债总额/GDP	8
国际储备	6	国际国内存款利率差	7
经常项目差额/GDP	5	实际汇率	5
偿债率	6	汇率波动幅度	10
M2/国际储备	7	货币供应量	6
国际信贷增长率	8	中央政府财政赤字	7

依照算法分析中的方法，分别取 $CSup^k_{min}=5\%$、$CConf^k_{min}=1.7\%$ 和 $\lambda=0.8$，

$$Sup^k_{min}=(\frac{k-1}{n-1})^2(CSup^n_{min}-Sup^1_{min})+Sup^1_{min},$$

其中 Sup^1_{min} 为一维的最小支持度，根据用户的偏好来选择。Sup^k_{min} 的衰减曲线见图1。

候选集的聚类组合数目随维数的变化见图2。挖掘到的模糊规则数目随 Sup^k_{min} 和 $Conf^k_{min}$ 变化曲线见图3。挖掘到的部分单指标模糊规则形式见表2。

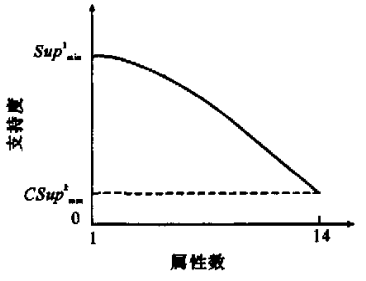

图1 支持度变化($Sup^k_{min}=10\%$ 和 $Conf^k_{min}=10\%$)

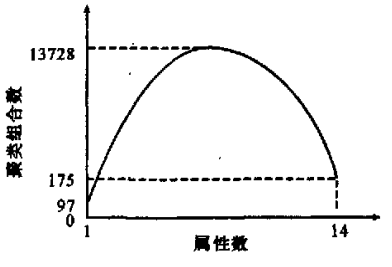

图2 候选集中聚类组合数目变化($Sup^k_{min}=10\%$ 和 $Conf^k_{min}=10\%$)

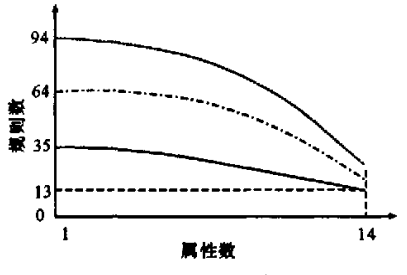

图3 规则数目变化

……：$Sup^k_{min}=5\%$ 和 $Conf^k_{min}=5\%$；－·－：$Sup^k_{min}=8\%$ 和 $Conf^k_{min}=8\%$；——：$Sup^k_{min}=10\%$ 和 $Conf^k_{min}=10\%$。

表2 单指标模糊聚类规则示例

序号	指标名称	区间		模糊规则集隶属度				
		下限	上限	不会发生	正常	可能发生	很可能发生	已经发生
1	GDP增长率	6	10000	0.259	0.436	0.305		
2	通货膨胀率	5	10		0.147	0.295	0.362	0.196
3	经常项目差额/GDP	−5	0		0.583			0.417

图1～3与表2的结果表明，随着维数的增加，指标间的相互影响和组合就更明显地显露出来了，这样可以更准确地来预测金融危机的发生概率。但是当维数很高时(如大于10)，规则数目反而降低，这主要是由于选择的指标间关联关系复杂，所有的指标同时恶化引发金融危机的情况较少，因此基于中低维模糊规则的金融危机预警就比较重要了。

4 结束语

本文通过对多维数据集中模糊规则的特性进行分析，提出多层次模糊规则的逐维算法。目前该算法已应用于金融危机预警系统中，在该系统中规则挖掘中的应用验证了算法的有效性，较好地处理了多属性的联合影响。

参考文献：

[1] JAY-LOUISE WELDON. Data mining and visualization[J]. Database Programming and Design, 1999, 21(5): 21-24.

[2] JUNE M, DONATO, JACK C, et al. Mining multi-dimensional data for decision support [J]. Future Generation Computer System, 1999(15): 433-441.

[3] 刘志强. 金融危机预警指标体系研究[J]. 世界经济, 2002, 23(4): 17-23.

[4] 范明，孟小峰. 数据挖掘[M]. 北京：机械工业出版社，2001: 223-254.

[5] HAN J. Mining knowledge at multiple concept levels [C]. //Proceedings of the 4th International Conference on Information and Knowledge Management (CIKM'95). Baltimore, Maryland, 1995: 19-24.

(责任编辑：邓大玉)