模糊概念格在知识发现的应用及一种构造算法

强 宇[1,2],刘宗田[1],林 炜[1],时百胜[1],李 云[1]
(1. 上海大学计算机学院,上海 200072;2. 蚌埠坦克学院,安徽蚌埠 233013)

摘 要: 基于有限 L-背景的模糊格在扩展和时空复杂度上有局限. 本文定义了广义的模糊概念格和其上的截运算以简化格构造,提出了一种模糊格构造算法. 在概念格结点级上定义了两个模糊参数 α 和 δ,以避免提取因高偏差导致的无效规则. 给出一个实例,说明了从模糊概念格提取不确定规则、计算规则支持度、置信度的原则、方法. 实现了构造算法与 Godin 算法的对比实验,结果表明本算法在时空性能上要优于 Godin 算法.

关键词: 人工智能; 属性模糊概念格; 模糊语言变量; 知识发现
中图分类号: TP18 **文献标识码**: A **文章编号**: 0372-2112 (2005) 02-0350-04

Research on Fuzzy Concept Lattice in Knowledge Discovery and a Construction Algorithm

QIANG Yu[1,2], LIU Zong-tian[1], LIN Wei[1], SHI Bai-sheng[1], LI Yun[1]
(1. *School of engineering and science in computer, shanghai university, Shanghai 200072, china;*
2. *BengBu Tank College, Bengbu, Anhui 233013, china*)

Abstract: Fuzzy lattice based on finite L - context has limit on expansion and time and spatial complexity. Fuzzy-attributes concept lattice in common sense and cut calculation on it were defined to simplify lattice construction. A construction algorithm was presented. Two fuzzy parameters α and δ on level of concept lattice node were defined to avoid extracting invalid association rules induced by high abbreviation. A case was given to demonstrate the principles for abstracting indefinite rules and computing their support and confidence. A comparison experiment between construction algorithm and Godin algorithm was made. The result shows that construction algorithm is superior to Godin algorithm on time-spatial complexity.

Key words: artificial intelligence; fuzzy attribute concept lattice; fuzzy language variable; knowledge discovery

1 引言

作为一种优良的形式分析工具,概念格已广泛应用于知识发现和数据挖掘. 从标准形式背景构造概念格的方法有渐进式构造[1]和批生成算法[1]等. 但实际中,信息往往模糊、不确定,故其表示及在知识发现中应用有深厚研究背景. Karl Erich Wolff[2]提出了一种模糊信息表示法,属性用模糊语言变量值表示,并构造语言变量值的分级格,用此格分类形式背景中对象. Burusco[3]从 L-模糊集合的形式背景,采用 fix point 理论,给出了一个计算格结构的方法. Ralam bondrainy 和 Girard[4,5]研究了由模糊量词描述的数据的概念格构造问题,通过应用语言变量代替二值的表示形式把"模糊"引入概念知识处理,通过在数据类型定义中插入模糊量词集处理模糊信息. 在数据类型中,精确数与模糊数并存. Karl Erich Wolff 的模糊信息表示将模糊语言变量的多值背景转换为单值背景,依据语言变量值的分级构造格,据分级格分类形式背景的对象. 此法有一定局限,从形式背景可直接构造概念格,而不需根据经验知识构造好属性分级格,再分类对象. 前者包含着聚类思想,更为自然,概念间的依赖关系更直观. 并且此表示是针对有限的 L-模糊背景做的,当 L-模糊集扩展到无限时,将无法表示. Burusco 给出的计算格方法也是针对有限 L - 模糊集的. 但实际模糊信息的取值多是连续的,故应考虑 L 无限的情况. Ralam bondrainy 关于模糊信息的处理是在数据类型定义中加入模糊量词,精确信息与模糊信息并存. 实际上精确信息取值可看成是模糊信息的特例. 故广义上,对象属性表可看成一个属性与对象的隶属度关系表,隶属度取一般函数. 文中第二部分给出了模糊形式背景及格的定义. 在格构造分析上,本文采用截运算处理模糊背景,去除小值隶属度,简化背景,提高格精确度,并在格结点级上定义了两个模糊参数 α 和 δ,使得从模糊格提取不确定规则时避免生成冗余的频繁结点对,进而避免生成因高偏差导致的无意义规则,从而使格构造与分析更简洁,快速.

收稿日期:2004-04-23;修回日期:2004-07-20
基金项目:国家自然科学基金《分布式概念格数学模型及算法研究》(No. 60275022)

2　属性模糊概念格

模糊集合是一种特殊定义的集合,隶属度函数反映了模糊集合中的元素属于该集合的程度.模糊集有很多表示法,但都须表示其包含的元素和相应隶属度函数.

定义 1　已知模糊形式背景为$(K:(O,D,I))$,O 为对象集,$O=\{o_1,o_2,\cdots,o_n\}$,D 为模糊属性标识集,$D=(d_1,d_2,\cdots,d_m)$,I 是函数,$O\times D\to[0,1]$,或写成 $0\leq m\leq 1$.隶属度函数可是一般函数.

定义 2　模糊形式背景的截运算为:对于每个 $d\in D$,定义 Φ_d,$0\leq\Phi_d\leq 1$,模糊形式背景的截运算为 $\lambda K=(O,D,I_\lambda)$,其中 O,D 同定义 1,I_λ 为:

$$I_\lambda(o,d)=\begin{cases} I(o,d),\text{如 } I(o,d)\geq\Phi_{d_i} \\ 0,\text{如 } I(o,d)<\Phi_{d_i} \end{cases}$$

定义 3　在 K 中,定义模糊概念 $c_i=(O_i,D_i)$,$O_i\subseteq O$,$D_i\subseteq D$,O_i 和 D_i 间可定义两映射 f 和 g,如下式表示:

$$\forall O_i\subseteq O: f(O_i)=\{d\mid \forall o\in O_i, I(o,d)>\Phi_d\}$$

$$\forall D_i\subseteq D: g(D_i)=\{o\mid \forall d\in D_i, I(o,d)>\Phi_d\},$$

f,g 称为 O 的幂集和 D 的幂集之间的 Galois 联接.

定义 4　如二元组(O_1,D_1) $(O_1\subseteq O, D_1\subseteq D)$ 满足:$O_1=g(D_1)$,$D_1=f(O_1)$,则称作 K 的一个模糊概念,O_1,D_1 是模糊概念 c_1 的外延和内涵.K 的所有模糊概念集合记为 $CS(K)$.$CS(K)$ 上的结构是通过泛化例化关系产生的,定义为:如 $O_1\subseteq O_2$ 则$(O_1,D_1)\leq(O_2,D_2)$.通过此关系得到有序集 $CS(K)=(CS(K),\leq)$,称作 K 的概念格.

生成概念格的方法有渐进式、批处理法、Godin 法等,渐进法经证明是优良的,故应用渐进的思想构造格.

3　属性模糊概念格的构造

基本算法思想:

(1)做数据预处理,对模糊属性-对象表中同一列的属性值计算 Φ_{d_i},小于 Φ_{d_i} 的属性值取 0.

(2)做数据预处理,将模糊属性-对象表中有相同属性集合(即对应相同的属性值均大于 Φ_{d_i})的对象合并为一个结点,得到处理背景.

(3)概念格初始化为空.生成根结点(Φ,D).

(4)从处理背景每加入一个对象 X^*,则生成新结点$(\{x^*\},f(\{x^*\}),\alpha,\delta)$,连接新点到根结点的边.

(5)从根开始,按自底向上,深度遍历方式与格中已有结点比较

(a)如格节点 c 内涵小于等于新增对象 x^* 的内涵(intent$(c)\subseteq f(\{x^*\})$),则格结点 c 更新点.更新为(extent$(c)\cup\{x^*\}$,intent(c),α,δ),边不更新.回到根.

(b)如格节点与新节点内涵有交集且不等于任格节点内涵,则向上搜索,找到与新节点有相同内涵交集的最大格节点(属性集相同,对象集最大),此为产生子结点,与新结点一起生成其父结点-新生节点(extent$(c)\cup\{x^*\}$,intent$(c)\cap f(x^*)$,α,δ).连接新生结点到其子结点的边.

(6)直到所有的对象加入格中.

(7)按自底向上方式搜索所有没有父结点的结点,生成顶结点(O,Φ),增加顶结点到这些点的边.

例子　在模糊逻辑中,模糊性语言称模糊语言,如长短、高矮等.模糊语言变量取值是用模糊语言表示的模糊集合.如年龄是模糊语言变量,则其值是"年轻"、"年幼"、"年老"等用模糊语言表示的模糊集合.对标准形式背景模糊化即是"模糊化"原有的对象和属性,用模糊隶属度函数值替代二元赋值.生成模糊背景如表 1:

表 1　模糊背景

	咳嗽	咳嗽	头疼	头疼	血压	血压	血压
对象	经常 $d1$	很少 $d2$	经常 $d3$	很少 $d4$	高 $d5$	中 $d6$	低 $d7$
1	**0.8**	0.2	**0.9**	0.1	**0.8**	0.2	0.0
2	**1.0**	0.0	0.0	**1.0**	**0.6**	0.4	0.0
3	**1.0**	0.0	0.1	**0.9**	**0.9**	0.1	0.0
4	0.3	**0.7**	**0.7**	0.3	0.0	0.6	0.4
5	0.6	**0.4**	**0.7**	0.3	0.0	**0.8**	**0.2**
Φ_{di}	0.74	0.26	0.48	0.52	0.46	0.42	0.12

在表 1 中,计算每列成员函数值的均值,做为阈值 Φ_{di}.超过此阈值 Φ_{di} 的取原值,粗体显示,加入概念生成.否则取 0.去除小值隶属度可提高构造格的精度,扫描模糊形式背景,将有相同属性集(即对应相同的属性值均大于 Φ_{di} 或均为 0)的对象合并,由此导出的处理背景如表 2 所示.

表 2　处理背景

	对象集	属性集	α	δ
a	1	$\{d_1,d_3,d_5\}$	0.83	0
b	2,3	$\{d_1,d_4,d_5\}$	0.9	0.1
d	4,5	$\{d_1,d_3,d_6,d_7\}$	0.56	0.11

对每个概念 $c(O_i,D_i)$ $(O_i\subseteq O, D_i\subseteq D)$,应用模糊背景的信息计算模糊参数 α 值和 δ 值,使从格提取关联规则时避免生成冗余频繁结点对,进而避免生成因高偏差导致的无意义规则,计算结果如表 2 和图 1.

α 定义为 c_i 中所有对象在所有属性中的隶属度的均值,反映了 c_i 的"平均模糊"程度,据模糊集合理论[6],先计算 c_i 的所有对象在模糊属性 d_i 上的隶属度均值 α_{d_i},因 c_i 包含属性集 D_i,故再计算 α_{d_i} 关于 D_i 的均值 α.

$$\alpha_{d_i}=\frac{1}{|O_i|}\sum_{o_i\in O_i} I(o_i,d_i)\ (|O_i|\text{ 是 }O_i\text{ 的元素个数})$$

$$\alpha=\frac{1}{|D_i|}\sum_{d_i\in D_i}\alpha_{d_i}\ (|D_i|\text{ 是 }D_i\text{ 的元素个数})$$

δ 定义为所有 $\delta_{d_i}(c_i)$ 的均值,其中 $\delta_{d_i}(c_i)$ 是所有对象在属性 d_i 中相对于 α 的均差,δ 反映了 c_i 中隶属度值的离散程度.根据模糊集理论[6],先计算 O_i 中所有对象在属性中 d_i 相对于 α 的方差 $\delta_{d_i}(c_i)$.因 c_i 包含属性集 D_i,故再计算 $\delta_d(O_i,D_i)$ 关于 D_i 的均值 δ.

$$\delta_{d_i}(O_i,D_i)=\delta(\{I(o_i,d_i)\mid O_i\in O_i\})=\sqrt{\frac{\sum_{o_i\in O_i}(I(o_i,d_i)-\alpha)^2}{|O_i|}}$$

$$\delta(O_i,D_i)=\frac{1}{|D_i|}\sum_{d_i\in D_i}\delta_{d_i}$$

应用渐进方法构造的模糊格如下:

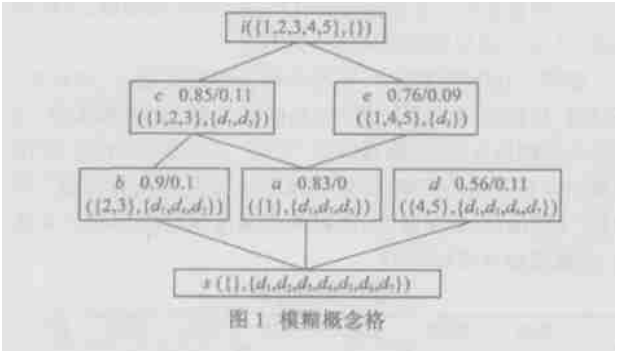

图 1 模糊概念格

4 由属性模糊概念格产生规则

关联规则是形如 $A \Rightarrow B$ 的表达式,A,B 为属性集,直观含义是:在数据库中具属性集 A 的对象可能也具备属性集 B.

关联规在数据挖掘中是重要方面,文[7]采用模糊集软化域的划分边界,且提出了模糊关联规则的概念,但未提出包含数据实际分布的划分算法及针对大型数据库的模糊关联规则挖掘算法. 文[8]采用语意云软化域的划分边界,提出了带语义云模型的关联规则挖掘算法. 文[9]采用关系模糊 C-平均算法将量化属性划分为几个模糊集,并提取关联规则.

在本文中,从模糊格挖掘关联规则的过程是:(a)找出所有频繁结点,(b)从频繁结点对提取关联规则. 根据模糊格可计算频繁结点.

给定背景中,标准偏差太大的概念应忽略,以避免抽取因高偏差导致的无意义规则.

产生规则的原则是:

(1)对模糊概念格中任意结点 c(外延集,属性集,参数 α,参数 δ),如 c 的 α 值大于某阈值 θ,δ 值小于某阈值 γ,则 c 是频繁结点.

(2)对模糊格中的频繁结点对 (c_1, c_2),如 c_1 和 c_2 为父子结点,且给定置信度阈值 Ψ,满足:$\frac{\min\{\alpha(c_1), \alpha(c_2)\}}{\max\{\alpha(c_1), \alpha(c_2)\}} \geq \Psi$,则 (c_1, c_2) 称 (θ, Ψ) 候选二元组.

(3)当且仅当 (c_1, c_2) 是 (θ, Ψ) 候选二元组,$A \Rightarrow B$ 是 (θ, Ψ) 关联规则,$A = \text{intent}(c_1)$,$B = \text{intent}(c_2) - \text{intent}(c_1)$;

例如设 α 的阈值 θ 为 0.6,δ 的阈值 γ 为 0.2,模糊格中 $\alpha > 0.6$,$\delta < 0.2$ 的结点均为频繁结点,图 2 频繁结点为 $a(\{1\}, \{d_1, d_3, d_5\})$,$b(\{2,3\}, \{d_1, d_4, d_5\})$,$e(\{1,4,5\}, \{d_3\})$,$c(\{1,2,3\}, \{d_1, d_5\})$.

设支持度阈值 $\theta = 0.6$,置信度阈值 $\Psi = 0.9$,则结点对 $\{e, a\}$,$\{c, a\}$,$\{c, b\}$ 均为候选二元组,可提取关联规则

表 3 由模糊概念格产生的部分规则及含义

序 号	规 则	解 释
1	$\{d_1\} \Rightarrow \{d_3, d_5\}$	如咳嗽经常,则头疼经常,血压高
2	$\{d_1, d_5\} \Rightarrow \{d_3\}$	如咳嗽经常,血压高,则头疼经常
3	$\{d_1, d_5\} \Rightarrow \{d_4\}$	如咳嗽经常,血压高,则头疼很少

关联规则 $A \Rightarrow B$ ($A = \text{intent}(c_1)$,$B = \text{intent}(c_2) - \text{intent}(c_1)$) 的支持度和置信度定义为:

$$\text{conf}(A \Rightarrow B) = \frac{\min\{\alpha(c_1), \alpha(c_2)\}}{\max\{\alpha(c_1), \alpha(c_2)\}}$$,阈值 Ψ.

5 实验结果及分析

为做实验评价,我们用 VC ++ 实现了该模糊格构造算法和 Godin 算法,并在 CPU 主频 PIV2.6G,内存 512K,操作系统 win2000 的 PC 机上做了对比实验. 测试数据采用 100 组随机生成的样本数据,为提高数据真实性,对随机生成的每 10 组样本数据取均值. 图 2,图 3 反映了两种算法空间复杂度的对比关系. 图 2 是属性集不变,集合数取 5,对象数变化与格规模的关系. 横坐标为对象数,从 20 始,每次增 20,到 200 止,纵坐标为初始格结点数,优化后对象数,优化后格结点数的取值. 显见,优化后对象数和格结点规模明显减少. 图 3 是对象集不变,集合数取 20,属性数变化与格规模的关系. 横坐标从 4 始,每次增 2,至 20 止,纵坐标为初始格结点数,优化后对象数,优化后格结点数的取值. 当属性数增加时,格结点数曲线变化大,可知,与对象数相比,本算法属性数变化对空间性能影响更大. 图 4、图 5 反映了两种算法的时间复杂度对比关系,图 4 为对象数固定(取 20),属性数变化与运行时间的关系曲线. 图 5 为属性数固定(取 5),对象数变化与运行时间的关系曲线. 图 5 曲线振幅大,可知,相对属性数改变,本算法对象数改变对时间性能影响更大. 据实验结果可见,此构造算法在空间性能上整体优于 Godin 算法.

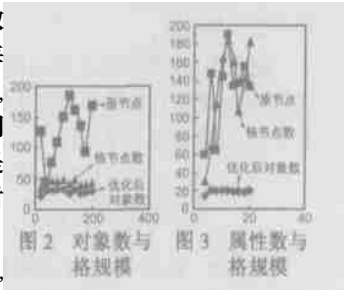

图 2 对象数与格规模 图 3 属性数与格规模

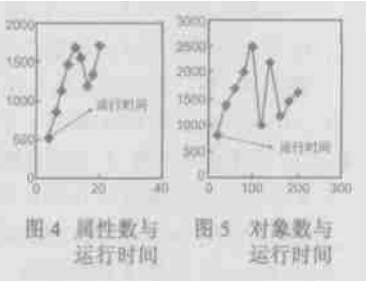

图 4 属性数与运行时间 图 5 对象数与运行时间

6 结语

本文阐述了将"模糊"引入概念知识系统,定义了广义属性模糊概念格和其上的截运算以简化格构造,并在概念格的结点级定义了两个模糊参数 α 和 δ 以避免提取因高偏差导致的无意义规则,使模糊格的构造和分析更简洁快速. 与多值背景的单值转换法相比,此法做了模糊表示,实际需要属性少,生成格规模小. 未来的研究方向还包括模糊格的快速生成算法、关联规则提取算法;大型格的分布式存储、及剪枝问题;格模型的扩展等.

参考文献:

[1] 谢志鹏. 概念格及扩展模型研究[D]. 合肥:合肥工大计算机学院,2000.12 - 20.

[2] Burusco A, Fuentes R. The study of L-fuzzy concept lattices[J]. Mathware Soft Comput, 1994, I(3): 209 - 218.

[3] Belohlavek R. Similarity relations in concept lattices[J]. Journal of Logic Computation,2000,10(6):823 - 845.

[4] Girard R,Ralambondrainy H. Conceptual classification from imprecise data[A]. Proceedings of Information Processing and Management of Uncertainty in Knowledge-Based System(IPMU'96)[C]. volume 1, Granada,Spain,1996. 247 - 252.

[5] Girard R,Ralambondrainy H. Conceptual classification from structured and fuzzy data[A]. Proceedings of the 6th IEEE International Conference on Fuzzy System[C]. Barcelona,Spain,1997. 135 - 142.

[6] 孙增圻,等. 智能控制理论与技术[M]. 北京:清华大学出版社,2000. 19 - 21.

[7] Chan M K,Ada F,Man H W. Mining fuzzy association rules in database[A]. Proceedings of the International Conference on information and Knowledge Managemnent[C]. LasVegas,Neveda,1997. 10 - 14.

[8] Lu Jianjiang,Qian Zuoping,Song Ziling. Application of normal cloud association rules on prediction[J]. Journal of Computer Research and Development,2000,37(11):1317 - 1320.

[9] Lu Jian jiang,Qian Zuoping,Song Ziling. Mining linguistic valued association rules[J]. Journal of software,2001,12(4):607 - 611.

作者简介:

强　宇　女,1971 年 4 月生于四川成都,上大计算机学院博士生,研究方向:数据挖掘. E-mail:qiangyu22@sina.com.cn.

刘宗田　男,1946 年 2 月生于山东淄博,上大计算机学院教授,博导,研究方向:人工智能,软件工程,高级编译技术.