

模糊概念格在知识发现中的应用研究^{*}

强 宇 刘宗田 吴 强 时百盛 李 云

(上海大学计算机学院 上海200072)

摘 要 概念格是一种优良的形式化分析工具,已经广泛应用于知识发现和数据挖掘中。在大量研究中概念格基于精确的形式背景,即二值背景,而在实际中,形式背景中的信息往往是模糊、不确定的。本文将“模糊”引入概念格,定义了属性模糊概念格和其上的截运算,在概念格结点上定义了两模糊参数 α 和 δ ,提出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,并给出了一个实例。

关键词 属性模糊概念格,模糊语言变量,知识发现

Research on Application of Fuzzy Concept Lattice in Knowledge Discovery

QIANG Yu LIU Zong-Tian WU Qiang SHI Bai-Sheng LI Yun

(School of Engineering and Science in Computer, Shanghai University, Shanghai 200072)

Abstract Concept lattice is a kind of excellent formal analyzed tool, has been widely applied in knowledge discovering and data mining. In many research Papers, concept lattice is based on precise context, also binary context. But in fact, information in context is usually fuzzy and indefinite. This paper introduces “fuzzy” into concept lattice, defines fuzzy-attributes concept lattice and cut calculation, defines two fuzzy parameters α and δ on level of concept lattice node, presents a method of abstracting association rules from fuzzy concept lattice and computing support and confidence of rules, also provides a case to illustrate the method.

Keywords Fuzzy attribute concept lattice, Fuzzy language variable, Knowledge discovery

1 引言

概念格是一种精确的数学模型,作为一种形式分析工具,已经广泛应用于知识发现和数据挖掘中。概念和属性间的关系构成概念并组成格形式,使得不同概念间的依赖变得更为明显。从已知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。Burusco 等在文[2]中从 L-模糊概念集的格结构,采用了 fix point 理论,给出了一个计算格结构的方法。文[3]从模糊集合理论中的三种不同蕴含算子出发,研究了与蕴含算子关联的模糊概念和格结构,该文说明了文[2]中的模糊概念格对应于 Kleene-Dienes 蕴含。文[4]研究了模糊概念格中的相似性关系,定义并研究了对象相似性(和属性相似性)、概念相似性、概念格相似性。还给出概念相似性简化概念格的方法。Girard^[5,6]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,据此思想,模糊概念格的形式背景承载的是模糊成员函数值而不是 true, false 值。用模糊语言变量表示标准属性,并保持属性与对象一致性,由此得到概念格是实数的。本文采用的方法是用截运算处理模糊背景,去除小值隶属度,提高格的精确度,并在格结点上定义了两模糊参数 α 和 δ ,使模糊格的构造与分析更简洁,快速。

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函数。可用序偶表示:

$$A = \{ (x, \mu_A(x)) | x \in X_i \}, X_i \text{ 为论域。}$$

可表示成更紧凑的形式

$$A = \left\{ \begin{array}{l} \frac{\mu_A(x)}{x}, X_i \text{ 连续} \\ \sum_{i=1}^n \frac{\mu_A(x_i)}{x_i}, X_i \text{ 离散} \end{array} \right\}$$

定义1 已知模糊形式背景为 $(K: (O, D, I))$, O 为对象集合; $O = \{o_1, o_2, \dots, o_n\}$, D 为模糊属性标识集合 $D = \{d_1, d_2, \dots, d_m\}$, I 是函数, $O \times D \rightarrow [0, 1]$, 或写为 $I(o, d) = m$, $0 \leq m \leq 1$ 。隶属度函数可是直线或曲线。

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

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

定义3 在模糊形式背景 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:

$$\forall O_i \subseteq O: f(O_i) = \{d_i | \forall o_i \in O_i, I(o_i, d_i) > \Phi_{d_i}\}$$

$$\forall D_i \subseteq D: g(D_i) = \{o_i | \forall d_i \in D_i, I(o_i, d_i) > \Phi_{d_i}\}$$

f, g 称 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: $O_1 = g(D_1)$, $D_1 = f(O_1)$, 则称为模糊背景 K 的一个模糊概念, O_1 ,

^{*} 本文受国家自然科学基金(分布式概念格数学模型及算法研究)(60275022)资助。强 宇 博士生, 讲师, 研究方向为人工智能的数据挖掘。刘宗田 博导, 教授, 研究方向为人工智能、软件工程、高级编译技术。吴 强 博士生, 研究方向为人工智能的数据挖掘。时百盛 博士生, 研究方向为数据挖掘。李 云 博士生。

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,其定义为:如 $O_1 \subseteq O_2$, 则 $(O_1, D_1) \leq (O_2, D_2)$ 。通过此关系得到的有序集 $CS(K) = (CS(K), \leq)$ 称作模糊形式背景 K 的概念格。

3 属性模糊概念格的构造

基本算法思想如下:

(0) 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 , 小于 Φ 的属性值取0, 忽略。

(1) 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) 的对象合并为一个结点。

(2) 概念格初始化为空。

(3) 从模糊属性-对象表每加入一个对象,设新增对象为 x^* , 则生成一个新结点,结点包括(外延,内涵,参数 α , 参数 δ)。

(4) 并与格中已有结点比较。

a) 如格节点 c 内涵小于新增对象 x^* 的内涵($\text{intent}(c) \subseteq f(\{x^*\})$), 则格节点 c 更新。 c 更新为 $(\text{extent}(c) \cup \{x^*\}, \text{intent}(c), \text{参数 } \alpha, \text{参数 } \delta)$ 。

b) 如格节点与新节点的内涵交集不等于任何格节点的内涵,按自底向上方式,找到与新节点有相同内涵交集的最大格节点(两节点内涵交集中对应的属性值均大于 Φ),此为产生子节点,与新节点一起生成一个新生节点。新生结点为 $(\text{extent}(c) \cup \{x^*\}, \text{intent}(c) \cap f(\{x^*\}), \text{参数 } \alpha, \text{参数 } \delta)$ 。

例1:初始形式背景如下表所示。

表1 初始形式背景

	咳嗽	头疼	血压
对象	次数	次数	
1	21	9	154
2	25	0	148
3	26	1	157
4	11	7	114
5	17	7	122

本文提出了采用模糊语言变量代替属性-对象的二元赋值的方法。在模糊逻辑中,模糊性的语言称模糊语言,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,则其取值是“年轻”、“年幼”、“年老”等用模糊语言表示的模糊集合。模糊隶属度函数表示了模糊集合中的元素属于该集合的程度。

对标准形式背景模糊化即是“模糊化”原有的对象和属性,即用模糊隶属度函数值替代二元赋值。

例如对于所给例1,三个模糊语言变量“咳嗽”、“头疼”和“血压”的值的集合分别为: $T(\text{咳嗽}) = \{\text{经常}, \text{很少}\}$, $T(\text{头疼}) = \{\text{经常}, \text{很少}\}$, $T(\text{血压}) = \{\text{高}, \text{中}, \text{低}\}$ 。上述每个模糊语言“经常”、“很少”是定义在论域上的模糊集合,用隶属度函数图可描述模糊函数,设隶属函数用三角函数表示,三个变量如图1所示。

根据隶属度函数处理初始背景,例如:

$$\mu_{\text{咳嗽经常}}(21) = 0.8 \quad \mu_{\text{咳嗽不经常}}(21) = 0.2$$

$$\mu_{\text{头疼经常}}(9) = 0.9 \quad \mu_{\text{头疼不经常}}(9) = 0.1$$

$$\mu_{\text{血压高}}(154) = 0.8 \quad \mu_{\text{血压中}} = 0.2$$

$$\mu_{\text{血压低}}(154) = 0;$$

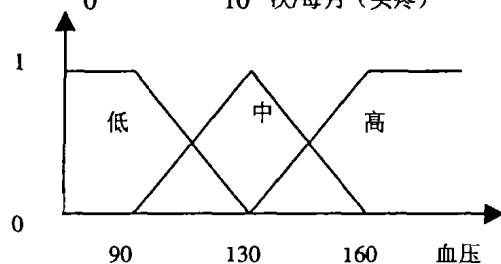
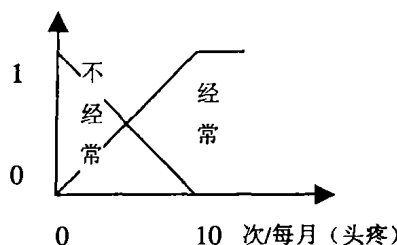
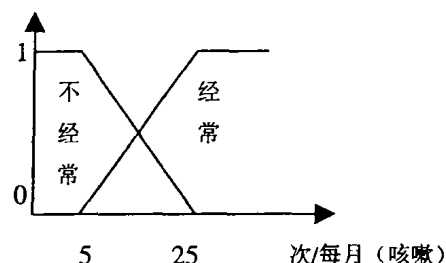


图1 隶属度函数图

可生成如下模糊属性-对象表。

表2 模糊属性-对象表

	咳嗽	咳嗽	头疼	头疼	血压	血压	血压
对象	经常 d_1	很少 d_2	经常 d_3	很少 d_4	高 d_5	中 d_6	低 d_7
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	0.74	0.26	0.48	0.52	0.46	0.42	0.12

经模糊化后,五个对象的属性对象表如表2所示,此法还可用于大型数据集的模糊化。在此模糊属性-对象表中,可计算成员函数值的均值,作为阈值 Φ 。超过此阈值 Φ 的取原值,可加入概念的生成,在表2中是粗体显示,否则取0,忽略。属性的平均值 Φ 表示对应于相应属性的对象的外延,对象具有的属性值超过均值的则粗体显示,并加入概念生成。如两对象有相同的属性集(即对应相同的属性值均大于 Φ),则两对象属于同一概念。由此导出的概念如表3所示。

对每个概念 $c(O_i, D_i)$ ($O_i \subseteq O, D_i \subseteq D$), 计算模糊参数 α 值和 δ 值,计算结果如表3和图2所示。

α 表示 c_i 中所有对象在所有属性中的隶属度的均值。反映了模糊概念 c_i 的“平均模糊程度”,据模糊集合理论^[7],先计算 c_i 的对象集 O_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 。

$$\alpha_{d_i} = \frac{1}{|O_i|} \sum_{o_i \in O_i} I(o_i, d_i)$$

$$\alpha = \frac{1}{|D_i|} \sum_{d_i \in D_i} \alpha_{d_i}$$

($|O_i|, |D_i|$ 是集合 O_i, D_i 的元素个数)

δ 是所有 $\delta_{\phi}(c_i)$ 的均值,据模糊集理论^[7],先计算 O_i 中所有对象在属性 d_i 中相对于 α 的方差 $\delta_{\phi}(c_i)$ 。因 c_i 包含属性集 D_i ,故再计算 $\delta_{\phi}(O_i, D_i)$ 关于 D_i 的均值 δ 。

$$\delta_d(O, D) = \delta(\{I(o, d) | o \in O\})$$

$$= \sqrt{\frac{\sum_{o \in O} (I(o, d) - \alpha)^2}{|O|}}$$

$$\bar{\delta}(O, D) = \frac{1}{|D|} \sum_{d \in D} \delta_d$$

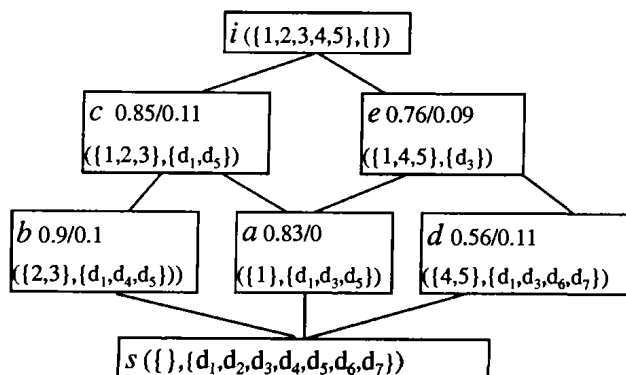


图2 模糊概念格

此概念格是层次结构的,越往上层去,对象集合越大,概念越粗化,沿着格从上到下的单链,对象集缩小,概念细化。

4 由属性模糊概念格产生规则

从结果概念格可提取关联规则,关联规则在知识库发现中是有价值的规律,是形如 $A \Rightarrow B$ 的表达式, A, B 为属性集,其直观含义是:在数据库中具备属性集 A 的对象可能也具备属性集 B 。

关联规则以往在数据挖掘中是一个重要方面,其中有 Ref^[8]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,也未给出针对大型数据库的模糊关联规则的挖掘算法。Ref^[9]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。Ref^[10]采用关系模型 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。本文构造的是属性模糊概念格,故提取的是模型规则。

关联规则的挖掘过程是:一、找出所有频繁结点,二、从频繁结点对提取关联规则。根据模糊格可计算频繁结点。

在给定的形式背景中,标准偏差太大的那些概念应忽略,使之不参与新的属性蕴涵式的推导过程,以避免抽取因高标准偏差导致的无意义蕴涵式。

表3 模糊概念的属性集-对象集

	属性集	对象集	α	δ
a	咳嗽经常,头疼经常,血压高	1	0.83	0
b	咳嗽经常,头疼很少,血压高	2,3	0.9	0.1
c	咳嗽经常,血压高	1,2,3	0.85	0.11
d	咳嗽很少,头疼经常,血压正常/低	4,5	0.56	0.11
e	头疼经常	1,4,5	0.77	0.09
s	所有属性	无对象	—	—
i	无属性	所有对象	—	—

下面给出几条产生规则的原则:

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. 当且仅当频繁结点对 (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 $a(\{1\}, \{d_1, d_3, d_5\})$, $b(\{2,3\}, \{d_1, d_4, d_5\})$, $e(\{1,4,5\}, \{d_1\})$, $c(\{1,2,3\}, \{d_1, d_5\})$ 。

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: $\{d_1\} \Rightarrow \{d_3, d_5\}$, $\{d_1, d_5\} \Rightarrow \{d_3\}$, $\{d_1, d_5\} \Rightarrow \{d_4\}$ 。

表4 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 $A \Rightarrow B$ ($A = \text{intent}(c_1)$, $B = \text{intent}(c_2) - \text{intent}(c_1)$) 的支持度和置信度如下定义:

$\text{sup}(A \Rightarrow B) = \alpha(c_2)$, $\alpha(c_2)$ 为结点 c_2 的 α 值;阈值为 θ ,

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

阈值为 ψ 。

结论 本文阐述了将“模糊”引入概念知识系统,定义了属性模糊概念格和其上的截运算,提出了从模糊格提取不确定性规则、计算规则支持度、置信度的方法。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,生成格的规模小。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

- 谢志鹏. 概念格及扩展模型研究: [合肥工业大学博士论文]. 2000
- Burusco A, Fuentes R. The study of L-fuzzy concept lattices. Mathware Soft Comput, 1994, 1(3): 209~218
- Burusco A, Fuentes-Gonzalez. Concept lattices defined from implication operators. Fuzzy Sets and Systems, 2000, 114: 431~436
- Belohlavek R. similarity relations in concept lattices. Journal of Logic Computation, 2000, 10(6): 823~845
- Girard R, Ralambondrainy H. Conceptual classification from imprecise data. In: Proc. of Information Processing and Management of Uncertainty in Knowledge-Based System (IPMU'96), volume 1, Granada, Spain. 247~252
- Girard R, Ralambondrainy H. Conceptual classification from structured and fuzzy data. In: Proc. of the 6th IEEE Intl. Conf. on Fuzzy System, Barcelona, Spain, 1997. 135~142
- 孙增圻, 等. 智能控制理论与技术. 清华大学出版社, 19~21
- Chan M K, Ada F, Man H W. Mining fuzzy association rules in database [A]. In: Proc. of the Intl. Conf. on information and Knowledge Management [C], Las Vegas, Nevada, 1997. 10~14
- 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
- Lu Jianjiang, Qian Zuoping, Song Ziling. Mining linguistic Valued association rules [J]. Journal of software, 2001, 12(4): 607~611