

粗集运算的产生式关联规则获取方法

陈功平, 秦乐阳

(信阳师范学院 网络信息与计算中心, 河南 信阳 464000)

摘 要: 分析了粗集运算的产生式关联规则(知识)获取方法的原理, 给出了其关联规则自动获取过程. 通过隶属函数的分类定义并求解其分类隶属度, 其目的是对数值型数据进行定性化处理, 从而形成规范的决策表. 研究了数据元组范化的方法、属性归约的具体算法. 提出了用多维代表值的求解方法进行元组值约简, 并给出了其具体算法. 实例表明, 通过该方法抽取的关联规则真实、可靠, 有效地实现了关联规则的自动获取且其操作的自动化程度高.

关键词: 粗糙集运算; 产生式关联规则; 属性归约; 多维代表值

中图分类号: TP182

文献标识码: A

文章编号: 1000-7180(2011)05-0113-05

Acquisition Method about the Production Association Rule Based on Rough Set Operation

CHEN Gong-ping, QIN Le-yang

(Network Information Center, Xinyang Normal University, Xinyang 464000, China)

Abstract: The basic principles of production association rule in rough set operation are analyzed first, and then the automatic process of production association rule acquisition is given. The numerical data is qualitatively processed by the classification of membership functions and membership degrees to get the normative decision table. The regular method of relations and the reduction algorithm of attributes are studied. The reduced relations are presented by the multi-represent-value method and its algorithm is offered. The whole production association rule acquisition process has high degree of automation and the extracted knowledge is true and reliable.

Key words: rough set operation; production association rule; attribute reduction; multi-dimensional representative value

1 引言

知识的自动获取一直是人工智能研究的难点, 也是专家系统开发人员一直追求解决的问题. 目前也有许多学者采用粗集运算的方法对数据进行处理, 最终提炼出强关联规则, 但这些方法存在操作运算复杂、整个操作自动化程度不高以及具体实现比较困难等缺点.

文中提出了一种基于粗集运算的产生式强关联规则获取方法的特点是: 整个知识获取过程的各个步骤均便于程序的自动化实现, 通过该方法抽取的知识真实、可靠. 该方法在一定程度和范围上解决了

信息系统知识自动获取的“瓶颈”问题, 为建造专家系统开发工具提供多种知识获取方法具有重要的现实意义.

2 关联规则自动获取原理

在产生式知识规则的模式下, 通过粗集理论的基本集合运算法则, 对大量的数据样本通过划分论域来确定隶属函数. 在求解隶属度的基础上对数值型数据进行定性化处理, 得出规范的 ACCESS 数据库决策表. 对规范的决策表使用属性迁移和概念树提升技术, 逐个属性地进行泛化, 最后得到简化的泛化关系. 对集合属性的辨识矩阵及辨识函数的计算

收稿日期: 2010-07-10; **修回日期:** 2010-08-16

基金项目: 河南省自然科学基金项目(112300410234)

找出规则前件的核属性,通过对属性约简集中的非核属性的可靠度计算,删除不必要的属性;通过对决策表的多维代表值的求解找出各元组的多维代表值,从而完成对决策表元组的约简(值约简).提取规范的产生式规则,计算其规则的各前提条件对结论的依赖度及整条规则的可信度,并对规则中的数据

进行反模糊化处理,从而得出自然语言描述的产生式规则.整个过程的每一个环节均可通过 VC++ 环境下的 ADO 技术的调用自动实现.

3 粗集运算的关联规则自动获取过程

粗集运算的关联规则自动获取过程如图 1 所示.

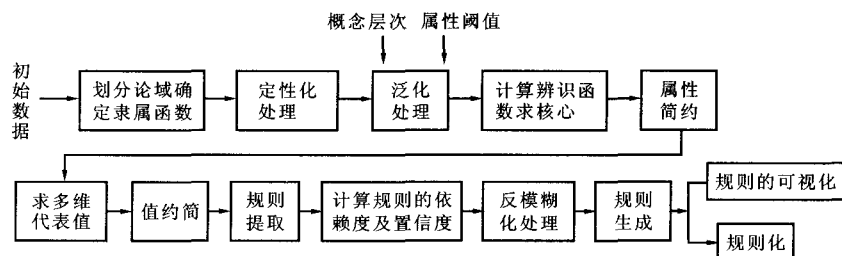


图1 粗集运算的关联规则自动获取过程

3.1 数据的定性化处理

专家系统知识获取方法之一是从数据样本中抽取出数据间的逻辑关系,最终提炼成产生式规则.而对大量数值型决策元组进行粗集处理,必须进行数据的定性化处理.定性化的过程就是对数值型数据进行划分分类,用相应的语言值代替数值型数据.文中采用黄金分割模型对数值数据进行定性化处理.将决策表各属性 a 的值划分为奇数个语言值,越靠近论域中心的其覆盖范围越小,相邻的语言值,离中心较近的语言值的熵是较远的 0.618 倍^[1].假设 $x_{\min}$ 、 $x_{\max}$ 分别是论域的最小值和最大值,取 5 个语言值,用钟型隶属度函数表示,从左到右分别为: $G_{-2}(\mu_{-2}, \sigma_{-2})$ 、 $G_{-1}(\mu_{-1}, \sigma_{-1})$ 、 $G_0(\mu_0, \sigma_0)$ 、 $G_{+1}(\mu_{+1}, \sigma_{+1})$ 、 $G_{+2}(\mu_{+2}, \sigma_{+2})$. 其均值和均方差如下:

$$\begin{cases} \mu_{-2} = x_{\min} \\ \mu_{-1} = \mu_0 - 0.382 \times (x_{\max} - x_{\min})/2 \\ \mu_0 = (x_{\max} + x_{\min})/2 \\ \mu_{+1} = \mu_0 + 0.382 \times (x_{\max} - x_{\min})/2 \\ \mu_{+2} = x_{\max} \\ \sigma_{-2} = \sigma_{-1}/0.618 \\ \sigma_{-1} = 0.318 \times (x_{\max} - x_{\min})/6 \\ \sigma_0 = 0.618 \times \sigma_{-1} \\ \sigma_{+1} = \sigma_{-1} \\ \sigma_{+2} = \sigma_{-2} \end{cases} \quad (1)$$

将数值型数据代入函数,分别得到各个语言值的隶属度,以隶属度最大的那个语言值代替数值数据,得到定性化后的决策表.

3.2 数据规范化

为了提高属性归约的效率,在使用归约算法前

先将初始数据泛化,缩减或浓缩决策表中的元组.也就是使用属性迁移和概念树提升技术,逐个属性地进行泛化,最后得到一个简化的泛化关系^[2].

下列算法用于如何从一个初始信息系统生成一个泛化的信息系统,其中, d_i 表示关系 R 中某个属性 a_i ($i = 1, 2, \dots, n$) 的不同值的个数, t_i 表示“属性阈值”, p 表示支持泛化的全部元组在整个决策表中所占的百分比(支持率),泛化时由用户指定.

算法 从某个关系中抽取泛化的信息系统.

- ① $\max \text{ tuples} \leftarrow p \times |R|$; $R' \leftarrow R$;
- ② 计算每个属性的 d_i 值;
- ③ While $|R'| \geq \max \text{ tuples}$ and $\exists d_i > t_i$
 {选择属性 a_i , 使 $a_i \in A$, 并满足 d_i/t_i 最大
 If 属性 a_i 是可泛化的

Then 在 H_i 中将 a_i 的概念提升一级,并在 R' 中对 a_i 下的概念进行替换

Else 从 R' 中删除属性 a_i

End If

删除 R' 中重复元组,重新计算 d_i 值}

End while

经过上述方法对数据集的处理可以大大提高数据归约的效率.

3.3 属性归约

3.3.1 通过辨识矩阵及辨识函数求属性的核心

属性的核心可作为计算归约集的基础,它是由辨识矩阵内仅含有单个元素构成的集合组成^[3].

令 $S = (U, A, V, f)$ 是一个知识表达系统, $A = C \cup D$, $|U| = n$, S 的辨识矩阵 M 是一个 $n \times n$ 矩阵,其任一个元素为:

$$\alpha(x_i, x_j) = \begin{cases} \Phi & x_i, x_j \in D \text{ 的同一个等价类} \\ \{a \in A \mid f(x_i, a) \neq f(x_j, a)\} & \\ x_i, x_j \in D \text{ 的不同等价类} & \end{cases} \quad (2)$$

因此, $\alpha(x_i, x_j)$ 是区别对象 x_i 和 x_j 的所有属性的集合.

为了求得辨识矩阵中的核,引入一个布尔函数,称其为辨识函数,用 Δ 表示. 对每个属性 $a \in A$, 指定一个布尔变量“ a ”. 若 $\alpha(x_i, x_j) = \{a_1, a_2, \dots, a_k\} \neq \Phi$, 则指定一个布尔函数 $a_1 \vee a_2 \vee \dots \vee a_k$, 用 $\sum \alpha(x_i, x_j)$ 来表示. 若 $\alpha(x_i, x_j) = \Phi$, 则指定布尔常量为 1. 辨识函数 Δ 可定义为:

$$\Delta = \prod_{(x_i, x_j) \in U \times U} \sum \alpha(x_i, x_j) \quad (3)$$

由式(3)知,利用布尔代数的分配律、结合律及吸收律求得该函数的极小析取范式的所有合取式,其极小范式即为属性集 A 的所有约简.

约简集是满足能区别由整个属性集区别的所有对象的属性极小子集. 故如果 $B \subseteq A$ 满足条件 $B \cap \alpha(x_i, x_j) \neq \Phi, \forall \alpha(x_i, x_j) \neq \Phi$ 的极小子集(关于包含),则 B 是 A 的一个约简.

而核是辨识矩阵中所有单个元素组成的集合,即:

$$\text{core}(A) = \{a \in A \mid \alpha(x_i, x_j) = \{a\}, x_i, x_j \in U\} \quad (4)$$

根据下面的定理,可以通过辨识矩阵计算出核心^[3].

定理 设有信息系统 $S = \{U, A, V, f\}$, $A = C \cup D$. 对于任何 $c \in C, c \in \text{core}(C, D), M(C) = (m_{ij})$, 当且仅当存在 $1 \leq j < i \leq n, m_{ij} = \{c\}$. 即 $\text{core} = \{c \in C \mid m_{ij} = \{c\}, \text{对于所有 } 1 \leq j < i \leq n\}$.

3.3.2 最简属性归约集的求解

在决策过程中,每个归约集都可以代替整个条件属性集,而不改变原有的依赖关系. 通常情况下,一个信息系统可能有多个属性归约集^[4]. 但每一个归约集必含有属性核,但除核属性之外的属性如何删除,可采用的方法为:先求得等价类的依赖度,然后计算除去约简核属性之外的属性的依赖度,从而作出是否删除的判断.

在属性归约中,利用两个属性集合 $P, R \subseteq A$ 之间的相互依赖程度,可以确定某一属性 a 的重要性. 属性集 P 对 R 的依赖程度用 $\gamma_R(P)$ 表示. 定义为:

$$\gamma_R(P) = \frac{\text{card}(\text{POS}_R(P))}{\text{card}(U)} \quad (5)$$

$$\text{POS}_R(P) = \bigcup_{X \in U/\text{IND}(P)} \text{apr}_R(X) \quad (6)$$

式中, $\text{card}(\cdot)$ 表示集合的基数; $\text{POS}_R(P)$ 是属性集 R 在 $U/\text{IND}(P)$ 中的正区域. 而对于任何一个属性集合 $P \subseteq A$, 不可分辨关系 IND 为:

$$\text{IND}(P) = \{(x_i, x_j) \in U \times U \mid f(x_i, a) = f(x_j, a), \forall a \in P\} \quad (7)$$

属性集 $C \subseteq A$ 对论域 U 进行划分得到的等价关系族记为 U/C .

对任何一个对象子集 $X \subseteq U$ 和属性子集 $R \subseteq A$, R 的下近似算子定义为:

$$\text{apr}_R(X) = \bigcup \{Y \in U/\text{IND}(R) \mid Y \subseteq X\} \quad (8)$$

通过属性依赖度的求解,然后计算约简集中某一属性的重要性来决定是否删除某一属性,考察条件属性 a 对于决策 d 的重要性,该指标可以用删除条件属性 a 前后的重要性之差 β 来表示^[5].

$$\beta = \gamma_R(P) - \gamma_{R-\{a\}}(P) \quad (9)$$

式中,若 $\beta = 0$,则表明条件属性 a 对于决策 d 没有影响,可以在系统中将其删除,经过分析处理之后,能够得到决策系统的约简.

具体属性约简算法描述如下^[6]:

令 M 是信息表 T 不考虑决策属性并删除了关于等价类的项之后所构成的辨识矩阵, M 的元素为 M_{ij} , $A = \{a_1, a_2, \dots, a_n\}$ 是 T 所有属性(不包括决策属性)的集合, $M_{ij} \subseteq A, a_k \in M_{ij}$.

计算各个条件属性 a_k 对决策 d 的重要性 β_k , 将 $\text{card}(M_{ij}) = 1$ 的所有属性 a_k 归入 C_0 , 令 $R = C_0$;

① $Q = \{M_{ij} \mid M_{ij} \cap R \neq \Phi, i \neq j, i, j = 1, 2, \dots, n\}, M = M - Q, B = A - R$;

② 对所有 $a_k \in B$, 计算其在 M 中的 β_k , 即 $\beta_k = \gamma_R(P) - \gamma_{R-\{a_k\}}(P)$, 并且将 $\beta_k > 0$ 时的 a_k 赋给 a_q , 即 $a_q = a_k$;

③ $R \leftarrow R \cup \{a_q\}$;

④ 重复 ① ~ ③, 直至 $M = \Phi$.

经过以上步骤得到的集合 R 是信息表 T 的一个属性约简.

3.4 值约简

对于二维表格的代表值约简算法,其算法主要基于如下思考:当一组同类元组中的一个或几个元组,其属性具有其它不同类元组所没有的值的时候,这些属性值是具有代表性的,即依靠这些值就可以识别这些元组,这些元组中的其他属性值则可以删

除. 称这些属性值为这些元组的代表值, 代表值是一个或几个属性值的组合, 代表值中属性的个数称为代表值的维数^[6]. 一维和二维代表值的寻找过程如下.

假设现有决策表 S , 它是经过上述各步聚的处理而得到的. 而决策表 S 通过变换得到新的二维表 T_1 , T_1 的行为所有涉及到约简的属性, 列为这些属性的取值, 其中 x 行 q 列的元素为在属性 x 上有相同属性值 q 的元组集合, 同类元组(决策值相同) 归入同一个集合中. 这样 T_1 的元素可能是多个集合. 对表 1 所呈现的信息表 S 转换后的 T_1 如表 2 所示.

表 1 信息表 S

U	a	b	c	D
x_1	0	0	-1	0
x_2	-1	0	-1	0
x_3	0	-1	0	-1
x_4	-1	-1	-1	-1
x_5	1	0	1	1
x_6	-1	-1	1	1
x_7	1	1	1	1

表 2 转化后的 T_1 表

	-1	0	1
a	$\{x_2, \{x_4\}, \{x_6\}$	$\{x_1\}, \{x_3\}$	$\{x_5, x_7\}$
b	$\{x_3, x_4\}, \{x_6\}$	$\{x_1, x_2\}, \{x_5\}$	$\{x_7\}$
c	$\{x_1, x_2\}, \{x_4\}$	$\{x_3\}$	$\{x_5, x_6, x_7\}$

从转换后的 T_1 表中取出仅有一类同类元组的元素, 即该元素中只有一类集合(单个集合), 但可能集合中有多个元素. 如表 2 中的 $\{x_5, x_7\}$ 、 $\{x_7\}$ 、 $\{x_3\}$ 、 $\{x_5, x_6, x_7\}$. 对属于同类元组的多个集合通过取“并”操作进行合并, 原则是尽可能的合并成最大的集合, 且合并后的集合是合并前的子集, 如 $\{x_5, x_7\} \cup \{x_7\} \cup \{x_5, x_6, x_7\} = \{x_5, x_6, x_7\}$. 经过第一次合并运算后的集合其所对应的属性值为一维代表值. 即合并运算后所得集合的行、列信息为该集合元素的一维代表值. 如表 2 中合并运算后集合 $\{x_5, x_6, x_7\}$ 中元组 x_5, x_6, x_7 的一维代表值为 $c = 1$, 元组 x_3 的一维代表值为 $c = 0$.

对表 2 中剩下的元素继续处理, 有多个同类元组的元素两两间进行“交”运算, 运算不可在同一属性间进行, 因为一个元组的属性不能同时有两个值, 如将 b_0 中的元素和 c_{-1} 中的元素进行交运算, 原则

是交运算后的集合也须是原集合的子集, 如 $(\{x_1, x_2\} \cup \{x_5\}) \cap (\{x_1, x_2\} \cup \{x_4\}) = \{x_1, x_2\}$, 依此类推, 最终可得到如表 3 所示的 T_2 表. 其中, 表 3 中的 b_0 表示属性 b 的取值为 0, 且表 3 为对称表.

表 3 转化后的 T_2 表

	a_{-1}	a_0	b_{-1}	b_0	c_{-1}
a_{-1}			$\{x_4\}, \{x_6\}$	$\{x_2\}$	$\{x_2\}, \{x_4\}$
a_0			$\{x_3\}$	$\{x_1\}$	$\{x_1\}$
b_{-1}					$\{x_4\}$
b_0					$\{x_1, x_2\}$
c_{-1}					

从上述处理后得到的 T_2 表中再次取出仅有一类同类元组的元素, 并对同类元素进行合并, 可以得到相应的二维代表值, 如 $\{x_2\} \cup \{x_1\} \cup \{x_1\} \cup \{x_1, x_2\} = \{x_1, x_2\}$, 即 $\{x_1, x_2\}$ 集合所对应的 b_0 、 c_{-1} 表明元组 x_1, x_2 的二维代表值为 $b = 0$ 且 $c = -1$; $\{x_3\}$ 集合所对应的 a_0, b_{-1} 表明元组 x_3 的二维代表值为 $a = 0$ 且 $b = -1$; 同理, 元组 x_4 的二维代表值为 $b = -1$ 且 $c = -1$. 然后检查所有元组是否通过上述运算要么有一维代表值, 要么有二维代表值, 或者具有两种代表值. 如某些元组不满足此条件, 可继续进行三维、四维运算. 但现实中基本上计算到一维、二维代表值即可.

借助这些一维和二维代表值, 可对简化后的决策表 S 进行属性值约简. 重复的元组予以合并, 具有多维代表值时, 因在属性归纳后的决策表中如果属性个数越多, 产生矛盾的可能性越小且问题描述的越真实, 故取高维代表值, 如元组 x_3 具有一维、二维代表值, 故取二维代表值. 从而得到如表 4 所示的信息表 S' (表 4 中“*”表示相应的属性值已删除).

表 4 信息表 S'

U'	a	b	c	D
x'_1	*	0	-1	0
x'_2	0	-1	*	-1
x'_3	*	-1	-1	-1
x'_4	*	*	1	1

3.5 关联规则提取

通过上述各步的处理, 最终得到合理的最简决策表, 以最简决策表中的条件属性作为前提条件的并, 而决策属性的值作为产生式规则的后件, 通过 VC++ 面向对象程序设计及 ADO 调用规范的决策

表产生出最终的产生式规则.如表4可得到值约简后的产生式规则如下:

R_1 : if $b = 0$ and $c = -1$ then $D = 0$

R_2 : if $a = 0$ and $b = -1$ then $D = -1$

R_3 : if $b = -1$ and $c = -1$ then $D = -1$

R_4 : if $c = 1$ then $D = 1$

令 R 表示产生式规则, $[x]_R$ 表示满足该规则条件属性的元组集合, D 表示满足该规则决策属性的元组集合, $\text{card}([x]_R)$ 表示元组集合中元组总数.令规则 R 的可信度为 Cf_k , 定义如下:

$$Cf_k = \frac{\text{card}([x]_R \cap D)}{\text{card}([x]_R)} \quad (10)$$

由于规则前件中每一个条件属性对得出该决策结论的贡献大小不同,也就是该前提属性对决策属性的依赖度不同,因此通过值约简后的决策表,重新计算没有删除前的条件属性对决策属性的依赖度 β , 通过式(9)计算出 β_{kj} ; 然后再将规则中简化的数值通过反模糊化处理,从而得到语言描述式的产生式知识规则.

至此已完成了规则的自动化抽取,从而得到可用的产生式规则^[7].

$$R_k : \text{AND}_{j=1}^m P_{kj}(\beta_{kj}) \rightarrow Q_k(Cf_k) \quad (11)$$

式中, P_{kj} 表示产生式规则的前提条件, Q_k 表示产生式规则的结论, k 表示新产生的规则序号, j 表示第 k 条产生式规则的第 j 个前提条件, m 表示该条规则的前提总数.

4 结束语

文中提出了基于粗集运算的产生式关联规则自

动获取知识的原理及过程.并提出了以多维代表值的求解以完成信息表的值约简;对获取知识过程中的关键步骤均给出了具体的算法.所获取的规则具有完整、准确、形式简洁、冗余度低,不必进一步化简等特点,知识获取的自动化程度高.系统可作为“知识自动获取机”嵌入到各种类型的专家系统中,具有较高的实用价值.

参考文献:

- [1] 韩敏,孙燕楠.地理信息知识获取 Rough—NN 模型研究[J].信息与控制,2005,34(1):104—108.
- [2] 刘同明.数据挖掘技术及其应用[M].北京:国防工业出版社,2001.
- [3] 王萍,王学峰.基于遗传算法的粗糙集属性约简算法[J].计算机应用与软件,2008,25(5):42—44.
- [4] 张文修,仇国芳.基于粗糙集的不确定决策[M].北京:清华大学出版社,2005:32—35.
- [5] 程玉胜,张佑生.粗集理论原型系统的集成与实现[J].微电子学与计算机,2006,23(3):119—121.
- [6] 谢孟军,黄国兴.基于 Rough Set 的规则自动抽取设计方案[J].计算机工程,2002,28(3):167—168.
- [7] 刘道华,原思聪.基于带区间值的产生式规则及其模糊推理机制研究[J].计算机应用与软件,2006,23(3):86—89.

作者简介:

陈功平 男,(1972—),硕士,讲师.研究方向为计算机网络、人工智能.

秦乐阳 男,(1982—),助理实验师.研究方向为计算机应用技术.

(上接第112页)

- [2] 高丹,彭新光.基于 DET 曲线的入侵检测评估方法[J].微电子学与计算机,2008(8):36—40.
- [3] Forrest S, Hofmeyr S, Somayia A. Computer immunology[J]. Communications of the ACM, 1997, 40(10): 88—96.
- [4] 余燕芳,陆军.基于改进遗传算法的服务器端负载均衡算法[J].微电子学与计算机,2007(7):41—45.
- [5] Hammad M, Conor R. A less destructive, context-aware crossover operator for GP[M]. Berlin, Heidelberg: Springer—Verlag, 2006:36—48.

作者简介:

盛云雷 男,(1985—),硕士研究生.研究方向为网络安全.

李永忠 男,(1961—),教授,硕士生导师.研究方向为计算机通信、信息安全、计算机应用.

荆春伟 男,(1984—),硕士研究生.研究方向为网络安全.