第33卷 第18期 计 算 机 工 程 2007年9月
Vol.33 No.18 Computer Engineering September 2007
·人工智能及识别技术· 文章编号：1000—3428(2007)18—0190—03 文献标识码：A 中图分类号：TP18

基于粗集理论的知识自动获取方法

刘道华[1,2]，原思聪[1]，李湘英[2]，王发展[1]
(1. 西安建筑科技大学机电学院，西安 710055；2. 信阳师范学院计算机科学系，信阳 464000)

摘 要：分析了粗集理论的知识自动获取方法的基本原理和获取过程，研究了数据记录范化方法、属性归约算法、最小决策规则集的求解、规则提取方法，并给出了自动获取方法的实例。实例证明了该算法的有效性。
关键词：粗糙集理论；知识自动获取；专家系统开发工具；属性归约；数据范化

Method of Automatic Knowledge Gain Based on Rough Sets Theory

LIU Dao-hua[1,2], YUAN Si-cong[1], LI Xiang-ying[2], WANG Fa-zhan[1]
(1. School of Mech. & Elec. Engineering, Xi'an University of Arch. & Technology, Xi'an 710055;
2. Dept. of Computer Science, Xinyang Normal University, Xinyang 464000)

【Abstract】 This paper analyzes the basic principle of automatic knowledge gaining method based on rough sets theory, introduces the automatic gaining process, and studies the method of data generalization, the concrete algorithm of the attribute reduction, solution of the smallest policy-making rule set and the method of decision-making rule. A concrete example is given about the knowledge's automatic gaining method. The example shows the method is effective and feasible.
【Key words】 rough sets theory; knowledge automatic gain; development tool of expert system; attribute reduction; data generalization

机械设计专家系统开发工具是当前机械设计制造业研究的热点，作为开发具体专家系统的母板，要提供多种知识自动获取方法，以适合不同知识表示及推理的知识自动获取的需要，知识的自动获取是人工智能知识自动获取研究的难点。在具体的应用规则库中，通过一定的算法获得新的规则知识，这是系统开发的难点。笔者通过数据范化算法，对规则库的记录集进行范化，通过具体知识规则库中的属性约简，在对规则库辨识矩阵及辨识函数的求解中，找出属性的核心，从决策矩阵及决策函数中提取最大泛化的规则(最小决策规则集)，从属性归约及数据范化后的规则库中，提炼新知识规则。具体实例证明该方法在一定程度上解决了信息系统知识自动获取的“瓶颈”问题。

1 知识自动获取原理

基于粗糙集理论的基本集合运算法则，通过对集合属性的辨识矩阵及辨识函数的计算，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，这样可大大简化规则的前件，而对规则的个数通过信息表的记录元组的不相关性(不可重复性)进行约简，通过对决策矩阵及决策函数的计算求解出最小决策规则集，并通过概念提升技术对约简后的信息表进行规则提取。整个过程的每一个环节均可通过编程来自动实现。

2 粗集理论的知识自动获取过程

对知识的数据进行预处理，并将其归一化为规范的数据库表，对该信息表通过辨识矩阵及辨识函数求出属性的核心，最终通过属性归约删除不重要的属性列，对删除属性列后的信息表剔除重复的元组，然后对元组再次进行数据范化，从而提取更一般的元组，最后通过计算决策矩阵及决策函数，求解出最小决策规则集，并从最小决策规则集中提取新知识规则存入知识库，从而完成新的知识获取过程，具体过程见图1。

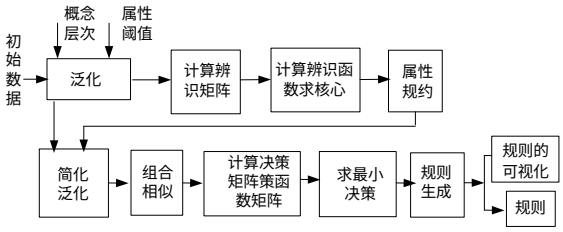

图1 粗糙集理论的知识自动获取过程

2.1 数据范化

实际数据库的元组数量往往很大，但有许多元组的信息是冗余的；有些元组数据虽然差异很少，但在可忽视的范围内；有的虽然不同，但代表的是同一类信息，可以等同视之。如果对这些数据不先进行处理或选择，而直接用粗糙集算法进行属性归约，效率势必很低。往往在使用归约算法前先将初始数据泛化，缩减或浓缩数据库中的元组。数据泛化的常用方法是：使用属性迁移和概念树提升技术，逐个属性地进

基金项目：国家自然科学基金资助项目(50275113)；陕西省教育厅产业化基金资助项目(04JC21)
作者简介：刘道华(1974－)，男，讲师、博士研究生，主研方向：智能系统，优化设计；原思聪，教授、博士生导师；李湘英，硕士、副教授；王发展，博士后、副教授
收稿日期：2006-11-21 **E-mail**：ldh30708@sina.com

行泛化，最后得到一个泛化关系[1]。

在基于粗糙集理论的数据归约中，元组归约和属性归约都是针对信息系统进行的，数据泛化的目的是生成泛化的信息系统，以便下一步针对该信息系统进行属性归约。

下列算法用于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 $a_i(i=1,2,...,n)$ 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数据库中所占的百分比(支持率)，泛化时由用户指定。如果一个属性是可泛化的(即存在相应的概念层次树 H_i)，并且 $d_i > t_i(t_i > 1)$，则称该属性相对于 t_i 是可泛化的。该算法的特点在于：用 d_i / t_i 值选择下一个被泛化属性，保证它具有最大 d_i / t_i 值，这样可以最大程度地改善泛化性能。

算法 从某个关系中抽取泛化的信息系统。

输入

(1)与任务相关的数据集 R，属性 $a_i(i=1,2,...,n)$；

(2)概念树集合 H，$H_i \in H$ 是可泛化属性 a_i 的概念层次；

(3)属性 a_i 对应的属性阈值 t_i，以及 $p(0 < p \le 1)$。

输出 泛化的信息系统 R'。

步骤

(1) $\max tuples \leftarrow p \times |R|$, $R' \leftarrow R$；

(2)计算每个属性的 d_i 值；

(3) *While* $|R'| \ge \max tuples$ *and* $\exists d_i > t_i$

Do {选择属性 a_i，使得 $a_i \in A$，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，

EndIf

删除 R' 中的重复元组，重新计算每个属性的 d_i 值}，

End while

利用上述方法对数据集进行处理，可以大大提高数据归约的效率。

2.2 属性归约

基于粗糙集理论的属性归约的一般步骤为：

(1)通过辨识矩阵及辨识函数求出属性归约集的核心；

(2)运用归约算法计算归约集，并根据归约集属性依赖度的计算删去不重要的属性，当数据量很大时，可在求核心前先将初始数据表泛化。

2.2.1 通过辨识矩阵和辨识函数求解属性的核心

属性的核心可作为计算归约集的基础，属性的核心是由辨识矩阵内仅含有单个元素构成的集合组成[2]。

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

$$\alpha(x_i,x_j)=\begin{cases}\Phi & x_i,x_j \in D\text{的同一等价类} \\ \{a \in A \mid f(x_i,a) \ne f(x_j,a)\} & x_i,x_j \in D\text{的不同等价类}\end{cases} \tag{1}$$

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(discernible function)，用 Δ 表示，对每个属性 $a \in A$，指定一个布尔变量“ a ”。

若 $\alpha(x_i,x_j)=\{a_1,a_2,...,a_k\} \ne \Phi$，则指定一个布尔函数 $a_1 \vee a_2 \vee \cdots \vee a_k$，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

$$\Delta = \prod_{(x_i,x_j) \in U \times U} \sum \alpha(x_i,x_j) \tag{2}$$

由式(2)可知，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”为属性集 A 的所有约简。

约简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。如果 $B \subseteq A$ 满足条件 $B \cap \alpha(x_i,x_j) \ne \Phi$，$\forall \alpha(x_i,x_j) \ne \Phi$ 的极小子集，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

$$core(A)=\{a \in A \mid \alpha(x_i,x_j)=\{a\}\} \tag{3}$$

其中，$x_i,x_j \in U$。

根据定理，可以通过辨识矩阵计算出核心[3]。

定理 设有信息系统，即

$S=\{U,Q,V,f\}$，$Q=C \cup D$

对于任何 $c \in C$，$c \in core(C,D)$，$M(C)=(m_{ij})$，当且仅当存在 $1 \le j < i \le n$，$m_{ij}=\{c\}$。即 $core=\{c \in C \mid m_{ij}=\{c\}, 1 \le j < i \le n\}$。

2.2.2 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 2 个属性集合 $P,R \subseteq Q$ 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

$$\gamma_R(P)=\frac{card(POS_R(P))}{card(U)} \tag{4}$$

$$POS_R(P)=\bigcup_{X \in U/IND(P)} \underline{apr}_R(X) \tag{5}$$

式(4)中，$card(\cdot)$ 表示集合的基数；式(5)中，$POS_R(P)$ 是属性集 R 在 $U/IND(P)$ 中的正区域。而对于任何一个属性集合 $P \subseteq Q$，不可分辨关系 IND 为

$$IND(P)=\{(x_i,x_j) \in U \times U \mid f(x_i,a)=f(x_j,a), \forall a \in P\} \tag{6}$$

而属性集 $C \subseteq A$ 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 $R \subseteq Q$，R 的“下近似”定义为

$$\underline{apr}_R(X)=\cup\{Y \in U/IND(R) \mid Y \subseteq X\} \tag{7}$$

通过属性依赖度的求解，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。

$$\beta=\gamma_R(P)-\gamma_{R-\{a\}}(P) \tag{8}$$

其中，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。

属性归约后并不改变原始属性集与决策属性之间的依赖程度，如果去掉该属性会造成依赖度变化，则恢复该属性，否则剔除该属性[5]。最后剩下的属性集就是最佳归约集或是用户定义的最小属性集。

2.3 最大泛化规则(最小决策规则集)的求解

对于决策属性 $d \in A$，及其特定值 V_d，关注的是满足 $d(e)=V_d$ 的对象 e 的集合 $\{V_d\}$。用矩阵形式可以将区分所有属于集合 $\{V_d\}$ 的对象和属于 $U-\{V_d\}$ 集合的对象的属性值对表示出来[6]。

(1)用 e_i 代表任何一个属于集合 $\{V_d\}$ 的对象， $i=1,2,\ldots$ ；$card(\{V_d\})=m$ ；

(2) e_j 代表任何一个属于集合 $U-\{V_d\}$ 的对象， $j=1,2,\ldots$ ；$card(U-\{V_d\})=n$。决策矩阵定义为

$$DM_{ij}=\{(a,a(e_i));a(e_i)\neq a(e_j)\} \tag{9}$$

其中，集合 DM_{ij} 包含所有 e_j 和 e_i 相异的属性值对。其形式上表示为矩阵 $\boldsymbol{DM}=[DM_{ij}]_{m\times n}$。

而从决策矩阵提取最大泛化规则(最小决策规则集)的基本方法为[7]

$$RUL=\bigcup_i RUL_i \tag{10}$$

其中，$RUL_i=\{r\in RUL\,|\,A_i(e_i)=(V_{i1},V_{i2},\ldots,V_{in})\}$ ；$e_i\in\{V_d\}$ ，符号 RUL_i 表示所有条件与对象 e_i 相匹配的最大泛化规则集。

对于给定的决策矩阵 $\boldsymbol{DM}$ 和决策值 V_d ，笛卡尔积 $\boldsymbol{F_i}=DM_{i1}\times DM_{i2}\times\cdots\times DM_{ir}$ 包含决策矩阵 $\boldsymbol{DM}$ 中第 i 行的所有序列对，由于 F_i 中的某些分量可能相同，因此在集合 $\tilde{\boldsymbol{F}}_{\boldsymbol{i}}=\{\{t\}\,|\,t\in F_i\}$ 中，$\{t\}$ 是向量 t 中所有不同分量的集合，$\tilde{\boldsymbol{F}}_{\boldsymbol{i}}$ 中的元素是对应于 $\{V_d\}$ 的所有规则，因此，这些规则至少匹配一个对象 $e_i\in\{V_d\}$ ，并且不与任何属于 $U-\{V_d\}$ 的对象相匹配。$\tilde{\boldsymbol{F}}_{\boldsymbol{i}}$ 中的规则是偏序的，其中最小的元素集合表示为 MIN_i。

MIN_i 集合的最大泛化规则可以通过一个布尔函数(决策函数)来计算。决策函数 B_i 由决策矩阵的第 i 行构造，即对于 $(DM_{i1},DM_{i2},\ldots,DM_{ir})$ 的分量，把出现在 DM_{ij} 中的每个属性值对作为一个布尔变量，并且对每个 DM_{ij} $(j=1,2,\ldots,n)$ 的分量进行“先析取后合取”的布尔运算，即

$$B_i=\bigwedge_j(\bigvee_i DM_{ij}) \tag{11}$$

通过式(11)可知，$\cup MIN_i$ 即为当前决策类的所有的最大泛化规则(最小决策规则集)。

2.4 规则提取

规则提取是通过实验或者从已经发生的事实中积累经验总结规则。应用粗集技术可以在大量的数据中，自动进行约简处理，进而得到简练的规则。这种方法弥补并解决了专家系统知识的“瓶颈“问题。根据粗集的约简原理，可以进行属性约简与值约简，约简的过程就是自动获取规则的过程。

3 应用实例

以某机组的工况为例，研究规则获取的步骤，转子碰磨故障示例见表 1。

表 1 记录信息

	a_1	a_2	a_3	a_4	D
x_1	0	1	0	0	0
x_2	1	0	1	0	0
x_3	0	1	1	1	1
x_4	1	0	0	1	0
x_5	1	1	0	1	1
x_6	1	0	1	1	1

表 1 中序号列表示各类对象， a_1 ～ a_4 列表示条件属性值， D 列表示决策属性值。由条件属性与决策属性来表示某一对象的特征，某一对象由条件属性进行分类。表内的数值 1 表示事件发生，数值 0 表示事件不发生。表 1 内 1～6 行表示测量的各次数据样本记录，a_1 ～ a_4 列表示属性：a_1 表示振动一倍频幅值，a_2 表示振动高频幅值，a_3 表示振动的相位变化， a_4 表示轴心轨迹的变化[8]。

通过式(1)，得到辨识矩阵 $\boldsymbol{M}$，即

$$\boldsymbol{M}=\begin{bmatrix}\Phi & & & & & \\ \{a_1,a_2,a_3\} & \Phi & & & & \\ \{a_3,a_4\} & \{a_1,a_2,a_4\} & \Phi & & & \\ \{a_1,a_2,a_4\} & \{a_3,a_4\} & \{a_1,a_2,a_3\} & \Phi & & \\ \{a_1,a_4\} & \{a_2,a_3,a_4\} & \{a_1,a_3\} & \{a_2\} & \Phi & \\ \{a_1,a_2,a_3,a_4\} & \{a_4\} & \{a_1,a_2\} & \{a_3\} & \{a_2,a_3\} & \Phi\end{bmatrix} \tag{12}$$

因为式(12)的辨识矩阵 $\boldsymbol{M}$ 是对称的，所以没有给出 $\boldsymbol{M}$ 矩阵的上三角元素。

由式(2)的辨识函数定义，可求得 Δ 为

$$\begin{aligned}\Delta=&(a_1\vee a_2\vee a_3)\wedge(a_3\vee a_4)\wedge(a_1\vee a_2\vee a_4)\wedge(a_1\vee a_4)\wedge(a_1\vee a_2\vee a_3\vee a_4)\\&\wedge(a_1\vee a_2\vee a_4)\wedge(a_3\vee a_4)\wedge(a_2\vee a_3\vee a_4)\wedge a_4\wedge(a_1\vee a_2\vee a_3)\\&\wedge(a_1\vee a_3)\wedge(a_1\vee a_2)\wedge a_2\wedge a_3\wedge(a_2\vee a_3)=a_2\wedge a_3\wedge a_4\end{aligned}$$

由上述辨识函数的计算结果可知，此知识表达系统有多个属性核，分别为 $\{a_2\},\{a_3\},\{a_4\}$ ，从该信息表的属性核求解看出，该信息表条件的最小属性集为 $\{a_2,a_3,a_4\}$。

令 $P=\{a_2,a_3,a_4\}$， $R=\{a_1,a_2,a_3,a_4\}$ ，通过式(4)、式(5)计算 $\gamma_R(P)$ ，通过式(8)计算删除条件属性 a_1 之后的重要性之差 β 可知， $\beta=0$ ，表明条件属性 a_1 对于决策 d 没有影响，可以从条件中删去。而删除属性 a_1 后，信息表的样本记录集可能有重复相同的元组，此时删除相同的样本记录集后组成表 2 的信息(注明本实例没有重复的记录集)。

表 2 记录信息

	a_2	a_3	a_4	D
x_1	1	0	0	0
X_2	0	1	0	0
x_3	1	1	1	1
x_4	0	0	1	0
x_5	1	0	1	1
x_6	0	1	1	1

通过概念提升技术及最小决策规则集的求解分析知：在样本记录 x_3 与 x_6 中 a_2 列的数据无论是 0 还是 1，对结论均无影响，可取样本记录 x_3 代替样本记录 x_6 的信息；样本记录 x_1 与样本记录 x_2 的属性 a_2 ， a_3 值为 0 或 1，对结果无影响，保留一行。结果记录见表 3。

表 3 最终的记录信息

	a_2	a_3	a_4	D
x_1	1	0	0	0
x_3	1	1	1	1
x_4	0	0	1	0
x_5	1	0	1	1

从表 3 中可以看到，样本记录 x_3 与 x_5 的属性 a_3 的值为 0 或 1 对结论无影响，因此二者的属性值计为“*”，样本记录 x_1 与 x_4 的属性 a_2 ， a_4 值对结果也无影响计为“*”，具体信息见表 4。笔者总结了 3 条规则，其中，“*”表示“发生”与否对结果无影响，即(1)若 a_2 , a_4 同时发生，则碰磨发生(从样本记录 x_3 和 x_5 看出)；(2)若 a_4 不发生，则碰磨不发生(从样本记录 x_1 看出)；(3)若 a_2 , a_3 不发生，则碰磨不发生(从样本记录 x_4 看出)。

表 4 记录信息

	A_2	a_3	a_4	D
x_1	*	0	0	0
x_3	1	*	1	1
x_4	0	0	*	0
x_5	1	*	1	1

4 结论

笔者给出了粗集理论知识自动获取的原理及知识自动获取过程，分析了系统泛化的具体算法，定义、求解了属性归

(下转第 207 页)