

## 基于粗糙集的约简算法及规则融合方法的研究

汪水友, 李 毅

(电子科技大学 计算机学院, 四川 成都 610054)

**摘 要:**从不同的角度分析了属性约简的两种重要方法:区分矩阵法和基于属性重要性。根据数据集的实际情况提出了一种基于粗糙集的区分矩阵和属性重要性相结合的启发式算法,并获得了属性约简集。在约简集的基础上分析了静态决策推理规则及算法。在相容决策系统中利用集合向量包含度构造了规则融合的方法,从而得到动态条件规则的极大近似决策值。在知识满足分类质量要求的前提下,根据规则融合方法,对任意给定的样本知识可以判别知识的实际归属类。

**关键词:**粗糙集;约简;规则;数据挖掘

**中图分类号:**TP18

**文献标识码:**A

**文章编号:**1673-629X(2008)07-0069-05

## Study on Rough Set Theory - Based Reduction Algorithm and Method of Rule Fusion

WANG Shui-you, LI Yi

(School of Computer Science & Engineering, University of Electronic  
Science and Technology of China, Chengdu 610054, China)

**Abstract:** Discernibility matrix and the significance of attributes were analyzed from different aspects, both of them were important methods of attribute reduction. A heuristic algorithm on rough set based on coalescent of discernibility matrix and the significance of attributes was raised. According to the algorithm a relative reduction set was formed. On the basis of the reduction set, the static decision rule reasoned and algorithm were analyzed. Finally the method of rule fusion was constructed by inclusion degree of set vector in the consistent decision system. The maximum approximate decision values were acquired from dynamic conditional rules. Under the condition of knowledge satisfied with classification quality, according to methods of rule fusion, its actual adscription class was found for any known sample knowledge.

**Key words:** rough set; reduction; rule; data mining

## 0 引言

粗糙集<sup>[1]</sup>(Rough Set)理论是一种新型的处理模糊和不确定知识的数学工具。它提供了一整套方法从数学上严格地处理数据分类问题,而且是处理具有信息不确定、不精确、不完善性系统的数学工具,是一种比较适用的归纳、分类方法。粗糙集理论仅仅分析隐藏在数据中的事实,并没有带人为的模糊性,是采用精确的数学方法分析不精确系统的一种理想方法。粗糙理论的基本思想是将数据库中的属性分为条件属性和决策属性,对数据库中的实例根据各个属性不同的属性值分成相应的子集,然后对条件属性划分的子集与决策属性划分的子集之间的上下近似关系生成判定

规则。

## 1 粗糙集理论基础

以下为粗糙集理论的相关定义<sup>[2]</sup>:

定义 1(信息系统) 信息系统是一有序对  $S = \langle U, A, V, f \rangle$ , 其中  $U$  为非空有限集合, 称为全域。全域  $U$  的元素被称为对象或者实例。 $A = C \cup D$ ,  $C$  是条件属性集合, 即对象的特征,  $D$  为决策属性集合, 称为对象的分类,  $C \cap D = \emptyset$ 。 $V$  是属性值的集合, 即  $V = \bigcup V_a$ ,  $V_a$  是属性  $a$  的值域;  $f: U \times A \rightarrow V$  是一个信息函数, 它为每个对象的每个属性赋予一个信息值, 即  $\forall a \in A, x \in U, f(x, a) \in V_a$ 。

定义 2(下近似)  $\underline{R}(X) = \{x \in U: [x]_R \subseteq X\}$  为概念  $X$  在关系  $R$  上的下近似。

定义 3(上近似)  $\overline{R}(X) = \{x \in U: [x]_R \cap X \neq \emptyset\}$  为概念  $X$  在关系  $R$  上的上近似。

收稿日期: 2007-10-07

作者简介: 汪水友(1974-), 男, 安徽合肥人, 硕士, 工程师, 研究方向为分布式数据库、数据挖掘; 李 毅, 博士, 教授, 研究方向为 OS 核心、机群技术、中间件。

概念  $X$  在关系  $R$  下的边界区域定义为:  $BN_R(X) = \bar{R}(X) - \underline{R}(X)$ , 边界区域就是在关系  $R$  下, 既可能属于  $X$  也可能不属于  $X$  的对象集合, 也是概念  $X$  的模糊性体现。

定义 4(近似精度) 等价关系  $R$  下, 概念  $X$  的近似精度为:

$\alpha_R(X) = |\underline{R}(X)| / |\bar{R}(X)|, X \neq \emptyset, |\underline{R}(X)|$  表示集合  $\underline{R}(X)$  的基数(元素个数)。

定义 5(属性依赖度) 在信息系统  $S = \langle U, A, V, f \rangle$  中,  $R, Q \subseteq A$ , 属性  $Q$  对属性  $R$  的依赖度定义为:

$k(R, Q) = \gamma R(Q) = |\text{POS}_R(Q)| / |U|$ ,  $R \Rightarrow kQ$  表示属性  $Q$  以  $k$  度依赖于属性  $R$  的。

当  $k = 1$  时,  $Q$  完全依赖  $P$ ; 当  $k = 0$  时,  $Q$  完全独立  $P$ ; 当  $0 < k < 1$  时,  $Q$  部分依赖  $P$ 。

定义 6(属性重要性) 在信息系统  $S = \langle U, A, V, f \rangle$  中, 设  $R, Q \subseteq A, a \in A, a \notin R, a \notin Q$ , 属性  $a$  相对于关系  $R$  对于  $Q$  的属性重要性定义为:  $\text{SGF}(a, R, Q) = k(R + \{a\}, Q) - k(R, Q)$ 。

定义 7(约简和核) 信息系统  $S = \langle U, A, V, f \rangle$  中,  $P \subseteq A, \text{IND}(P) = \text{IND}(A)$ , 称  $P$  是  $A$  的一个约简, 记为  $\text{RED}(A)$ 。  $\text{Core}(A) = \bigcap \text{RED}(A)$ 。即核为所有约简的交集。

定义 8(规则可信度)  $C = \{C_1, C_2, \dots, C_n\}, [X]_C$  是满足  $\text{Des}(c_1) \cap \text{Des}(c_2) \cap \dots \cap \text{Des}(c_n)$  的等价类,  $[d]_D$  是  $\text{Des}(d)$  的等价类, 那么规则的信任度为:

$$\gamma = |[X]_C \cap [X]_d| / |[X]_C|$$

定义 9(区分矩阵) 区分矩阵是一个对称  $|U| \times |U|$  矩阵, 任一元素:

$$a(x_i, x_j) = |a \in A \mid f(x_i, a) \neq f(x_j, a)|$$

$$\text{则区分函数 } \Delta = \prod \sum a(x, y)。$$

对于决策表区分函数要满足  $W(x, y)$ , 即  $x, y$  不能同时属于正域或者在同时属于正域的情况下不能同时属于决策表的等价类。

定义 10(包含度或粗糙隶属度)<sup>[3]</sup> 设  $U$  为有限论域, 对于任意  $X, Y \subseteq U$ , 称  $D(Y/X)$  为包含度,  $D(Y/X) = |X \cap Y| / |X|$ , 当  $|X| = 0$  时,  $D(Y/X) = 0$ 。

## 2 约简算法

首先阐述两种基本算法, 一种是用区分矩阵, 另一种是基于属性重要性。

(1) 用区分矩阵法约简, 首先构造区分矩阵, 在区分矩阵的基础上得出区分函数。然后应用吸收律对区

分函数进行化简, 使之成为析取范式, 则每个值蕴含式均为约简。基本算法可以求出所有的约简。实质上基于区分函数的方法就是将约简转化为对布尔函数的化简。

显然, 在区分矩阵中可以化简区分函数, 只有一个属性的项肯定是核, 含有核的析取式就可以置空, 也就是说可以对区分矩阵的析取表达式先行处理, 用一个一维数组  $\text{Arr}[n]$  来存取化简区分矩阵的每一行的所有元素, 从而去掉重复的元素; 这样就很容易找到约简表达式; 因为事先可以按照所有属性值的不同进行数据库的排序即降低了时间的复杂度。

计算等价类的时间复杂度为  $O(|A| \mid U \mid \log \mid U \mid)$ ,  $|A|$  为属性数,  $|U|$  为对象数。计算区分矩阵的时间复杂度也就是  $O(|A| \mid U \mid \log \mid U \mid)$ 。

(2) 基于属性重要性的启发式算法思想<sup>[4]</sup>: 以核作为计算约简的基础, 以属性的重要性作为启发信息, 按属性重要性的大小依次将属性加入约简集, 直到该集合是一个约简为止。按照加入属性的不同, 可以计算多个不同的约简集, 最终得到一个约简集。核的初始值可以为空, 但是指定的核还要用程序来判断, 当然经过区分矩阵得出是比较优化的。

该算法过程: 假设当前的候选约简集是  $P$ , 首先计算条件属性中不含  $P$  中的属性相对于决策属性  $D$  的重要度, 按照属性的重要程度从大到小逐个加入属性, 直至该集合是一个约简为止, 完成约简的扩充; 接着检查该集合中的每个非核属性, 删除该属性是否会改变该集合的对决策属性依赖度, 如果不影响, 则将其删除。其步骤是:

- 以信息表中的核和用户指定的属性集作为约简的初始值, 可以从空集开始计算;
- 分别计算  $\text{POS}_{\text{RED}}(D); \text{POS}_C(D)$ ;
- while  $\text{POS}_{\text{RED}}(D) \neq \text{POS}_C(D)$ ;
- 计算每个属性的  $\text{SGF}(a, \text{RED}, D)$  值, 在属性集取  $\text{MAX}(\text{SGF}(a, \text{RED}, D))$ , 则  $\text{RED} = \text{RED} + \{a\}$ ;
- 计算  $\text{POS}_{\text{RED}}(D)$ ;
- 对  $\text{RED}$  中的每个非核属性, 试去掉是否影响依赖系数; 如果不影响, 再删除此属性。

计算属性重要度的时间复杂度为:

$$O(|A|^2 \mid U \mid \log \mid U \mid)$$

(3) 在这里综合以上两种算法提出一种新的算法。算法分析:

区分矩阵的元素是区分各个不相同分类的条件属性集合, 那么要想把不同分类(决策属性值)区别开来, 靠的就是区分矩阵各元素。所以约简和区分矩阵元素的功能是相同的, 它们应该有交集<sup>[5]</sup>。

如果区分矩阵中只有一个属性,说明该属性一定是核集中的一员,也就是约简集中的一员。如果一个属性在区分矩阵中出现的频率越高,就可以近似理解成该属性对区分对象的重要程度就越大,就可以近似根据出现频率决定属性的重要程度。因此可以构造一个近似频率函数,而且考虑  $|a(x_i, x_j)|$  的值对函数值的影响更大,即令  $f(k) = |a(x_i, x_j)| + |A|/f(k)$ , 因为  $|a(x_i, x_j)|$  正好与属性在区分矩阵中出现的频率成反比,若  $|a(x_i, x_j)| = 1$ , 则它一定是核中的一个成员。

算法:求解约简集。

输入:一个条件属性集  $C$  和决策属性集  $D$  的协调信息系统  $S = \langle U, A, V, f \rangle$ 。

输出:一个约简集。

a. 初始化  $RED = \emptyset, f(a_i) = 1$ ;

b. 计算等价类  $U/A$ ;

根据  $A$  对对象进行排序(一般用快排),显然排序好的对象易于分割;

$k = 1, j = 1; A = \{x_1\}$ ;

For( $i = 2; i \leq |U|, i++$ )

If  $x_i$  和  $x_j$  对于  $A$  中的每个属性值都相同,则  $A_k = A_k \cup \{x_i\}$ ; // 合并等价类

Else  $k = k + 1; A_k = \{x_i\}; j = i$ ; // 建新的等价类

假设有  $N$  个等价类;  $N = 1, 2, \dots, |U|$ 。

c. 从每个等价类中取一对象来计算区分矩阵并同时计算属性的加权频率  $f(a_i)$ ;

d. 对区分矩阵按照  $f(a_i)$  进行降序排列;

e. for( $i = 1; i++; i \leq N$ )

for( $j = 1; j++; j \leq N$ ) // 遍历矩阵

f. for( $k = 1; k < |A|; k++$ ) // 遍历属性

g. if  $a(x_i, x_j) \cap RED = \emptyset$

选取  $\min(f(a_i))$ ;

h.  $RED = RED \cup \{a_i\}$ ;

}

i. return RED // 求得所有的约简

显然此算法较好利用了上述两个算法的优点,既考虑了属性的重要性,又利用了求核的矩阵特性,因此便于解决更多的实际问题。求等价关系的的最坏复杂度为  $O(|A| |U|^2)$ 。其中  $A$  为属性集合,  $U$  为对象集合。这是因为在最坏情况下需要扫描对象集合两次,每个对象一次,每个对象的等价类一次。文中改进的算法是首先按给定属性集对对象排序,然后扫描一遍即可。这样它的复杂度就降低到  $O(|A| |U| \log |U|)$ 。计算区分矩阵的时间复杂度为  $O(|A| |N|^2)$ 。对区分

矩阵排序的时间复杂度为  $O(2 |N|^2 \log |N|)$ , 后面在求  $\min(f(a_i))$  的过程中其时间复杂度  $O(|A| |N|^2)$ 。因此总的时间复杂度为:  $O(|A| |U| \log |U| + 2(|A| + 2 \log |N|) |N|^2)$ 。

### 3 决策规则和决策算法

$v_i \in v_{a_i}, P = \{a_1, a_1, \dots, a_n\}, P \subseteq A$ , 则称形如  $(a_1, v_1) \wedge (a_2, v_2) \wedge \dots \wedge (a_n, v_n)$  的公式为  $P$  基本公式, 简称  $P$  公式, 一个  $A$  基本公式称为一个基本公式。

对决策规则  $\phi \rightarrow \varphi$ , 如果  $\phi$  是一个  $P$  公式,  $\varphi$  是一个  $Q$  公式, 则称  $\phi \rightarrow \varphi$  是一个  $PQ$  基本决策规则。

表1为规则判断方法。

表1  $S = \langle U, C \cup D \rangle$

$U$	$a$	$b$	$c$	$d$	$e$
1	1	0	2	1	1
2	2	1	0	1	0
3	2	1	2	0	2
4	1	2	2	1	1

设  $P = \{a, b, c\}, Q = \{d, e\}$  分别是条件属性和决策属性, 则  $P, Q$  唯一地联系于如下的  $PQ$  决策算法:

$$a_1 b_0 c_2 \rightarrow d_1 e_1$$

$$a_2 b_1 c_0 \rightarrow d_1 e_0$$

$$a_2 b_1 c_2 \rightarrow d_0 e_2$$

$$a_1 b_2 c_2 \rightarrow d_1 e_1$$

显然以上都是静态确定性规则。

一个  $PQ$  决策算法中的规则  $\phi \rightarrow \varphi$  在  $S$  中是相容的当且仅当对任意该算法中的规则  $\phi' \rightarrow \varphi', \phi \rightarrow \phi'$  蕴涵  $\varphi \rightarrow \varphi'$ ; 从以上的定义知道  $\text{CORE}(P, Q) = \text{RED}(P, Q)$ ,  $\text{RED}(P, Q)$  表示算法  $(P, Q)$  的所有约简集合。

上面的决策样本分析都是静态的学习过程, 如果不断地增加新样本则是一种动态的学习过程。加入一个新样本有三种情况:

(1) 新样本与实际知识相同;

(2) 新样本与实际知识矛盾;

(3) 新样本完全是新情况。

为了表示新样本对已得出决策的影响, 可以计算上述情况对分类质量发生的变化。也就是根据样本的不同情况分别计算不同的分类质量。

如果新样本匹配现有的一类知识或完全是一类新知识, 则分类质量:

$$k = (r_c(D) = \text{card}(\text{pos}_c(D)) + 1) / (\text{card}(u) + 1)$$

若新样本与已获得的实际知识矛盾, 则分类质量:

$$k = (\text{card}(\text{pos}_C(D)) - \text{card}([x]_{\text{ind}(a)})) / \text{card}(u) + 1)$$

其中  $[x]$  表示新增的等价类。

其实还有一种情况就是新样本与实际知识只有少部分矛盾,可以认为在不影响判断决策的情况下近似认为这种样本属于已有的等价类中的一部分。用规则包含度可以近似地推出相应的决策规则,因此在下面将要讨论规则可信度的普遍规则。

#### 4 规则融合

信息融合利用已有的知识和经验来处理从未知世界得到的来自不同领域的的数据,是一个演绎的过程。规则融合也是用模型来融合数据,再给出所有不同条件下的相应决策。设  $(U, A, R, R_d)$  是协调近似空间,则  $R_A \subseteq R_d$ 。若  $B$  是属性约简集,则  $R_B \subseteq R_d$ 。令  $U/R_B = \{[x_i]_B \mid x_i \in U\}$ ,  $U/R_d = \{D_1, \dots, D_r\}$ 。由于  $R_B \subseteq R_d$ , 则对于任意  $x_i \in U$ , 必有  $D_j$  存在, 使  $[x_i]_B \subseteq D_j$ , 于是得到决策规则: 若  $x \in [x_i]_B$ , 则决策为  $D_j$ , 即有:

If  $\wedge (a_i, f_i(x_i))$ , then  $D_j$

简记为:  $(f_1(x_i), f_2(x_i), \dots, f_k(x_i)) \Rightarrow D_j$

于是对于每一类  $[x_i]_B$  可以得到一条确定性规则,显然有  $|U/R_B|$  条规则,也就是对属性约简和值约简后得到的静态规则。但是对于动态的新增数据在不同条件组合有  $|V|^m$  种情况(属性域为  $V$ )<sup>[6]</sup>。

现在要对任意给定的条件属性值能够推导最相近的决策值,那么首先要通过现有的知识建立一般概念下的模型,再用模型来推导所有的决策。当然有的决策可能不是其真实的结果,就用分类质量来反映其变化的程度。若偏离较大就要重建模型来决策。

其基本思想是通过等价分类分别形成条件等价类和决策等价类,并判别由条件属性类能否推导决策属性类,否则为不相容判别要另行处理<sup>[7]</sup>,若相容则根据已有的规则来分别判断新加进的知识属于哪一类。相应的步骤:

(1) 由决策属性  $R_d$  将  $U$  划分,得到决策等价类  $U/R_d = \{D_1, \dots, D_k\}$ , 并计算其正域  $\text{Pos}_C(D)$ , 判断是否为相容决策集。若不是相容决策集一般情况下用信息熵来进行推断近似的规则,当然在不保证绝对精度的情况下也可以将不相容决策集变为相容决策集,也就是在正域情况下的相容决策。

(2) 对于每一决策  $D_j (j \leq k)$  要搜寻相应的条件属性的值,形成基于决策类的集合向量。因为决策类是条件类的超集,其  $V_j = \{(\{f_1(x_i)\}, \dots, \{f_m(x_i)\}) \mid$

$[x_i]_A \subset D_j\}$ 。

(3) 将  $V_j$  中的对应向量取析取运算,即每个分量取析取运算,得到

$$F_j = (F_1^j, F_2^j, \dots, F_m^j) (j \leq m)$$

显然通过向量的合并,  $F$  中的向量组合显然比  $V$  中的向量要多,对于这些新组合的向量,其实都是间接隐藏在  $V$  中。因为一开始没有对属性进行约简,所以向量的合并与组合并没有降低判定规则的可信度。

(4) 对于任意  $v_l \in V_l (l \leq m)$ ,  $E = (\{v_1\}, \{v_2\}, \dots, \{v_m\})$ , 通过集合包含度(粗糙隶属度)公式计算每一个分量,再进行比较。

$$D(F_j/E) = |F_j \cap E| / |E|$$

假设  $D(F_j/E) = \text{MAX} D(F_j/E)$ , 则找到最佳匹配值。得到规则:

$$(f_1(x_i), f_2(x_i), \dots, f_k(x_i)) \Rightarrow D_j$$

(5) 对于任意给定的  $E$ , 从它的可信度来区分新增的知识属于哪一类,再根据分类质量  $k = (r_c(D))$  计算新知识对分类的影响。以其更好判断新知识对可信规则的影响程度,便于更合理地推导相应的决策值。

下面用决策表来计算可信度并推导其极大近似或确定性规则。如表 2 所示。

表 2  $S = \langle U, C \cup D \rangle$

$U$	$a_1$	$a_2$	$a_3$	$D$
$x_1$	2	5	3	1
$x_2$	3	2	1	2
$x_3$	2	5	3	1
$x_4$	2	2	3	2
$x_5$	1	1	3	3
$x_6$	1	1	2	3

由上面的规则判断方法来计算其等价类:

$$U/R_A = \{\{x_1, x_3\}, \{x_2\}, \{x_4\}, \{x_5\}, \{x_6\}\}$$

$$U/R_D = \{\{x_1, x_3\}, \{x_2, x_4\}, \{x_5, x_6\}\}$$

$$\text{显然 } \text{Pos}_C = \{x_1, x_3, x_2, x_4, x_5, x_6\} = U$$

即此表为协调决策表。

计算每一决策值对应的对象集合:

$$D_1 = \{x_1, x_3\}, D_2 = \{x_2, x_4\}, D_3 = \{x_5, x_6\}$$

计算在每一个决策值的条件等价类属性值:

$$V_1 = \{(\{2\}, \{5\}, \{3\})\}$$

$$V_2 = \{(\{3\}, \{2\}, \{1\}), (\{2\}, \{2\}, \{3\})\}$$

$$V_3 = \{(\{1\}, \{1\}, \{3\}), (\{1\}, \{1\}, \{2\})\}$$

析取集合的分向量:

$$F_1 = (\{2\}, \{5\}, \{3\})$$

$$F_2 = (\{3, 2\}, \{2\}, \{1, 3\})$$

$$F_3 = (\{1\}, \{1\}, \{2, 3\})$$

下面要根据具体的事例对三种情况进行说明:

1) 若给出  $E = (\{2\}, \{5\}, \{1\})$ , 按包含度公式分别计算  $E$  和  $F_1, F_2, F_3$  的近似度。

$$D(F_1/E) = 2/3, D(F_2/E) = 4/5, D(F_3/E) = 0$$

$F_2$  对应的包含度最大, 于是得到规则:

$$\text{If}(a_1, 2) \wedge (a_2, 5) \wedge (a_3, 1) \text{ then } d = 2 \text{ (0.8)}$$

显然这与已有的知识矛盾, 但是可信度大于 0.5 属于小部分矛盾, 可以近似推导。

2) 当  $E = (\{3\}, \{2\}, \{3\})$  时, 根据公式  $D(F_2/E) = 1$ , 而尽管  $(\{3\}, \{2\}, \{3\})$  不属于  $M$  中分量, 但是由于等价类中冗余的存在, 按照粗糙集原理同样可推导  $E = (\{3\}, \{2\}, \{3\})$  为确定性规则, 即

$$\text{If}(a_1, 3) \wedge (a_2, 2) \wedge (a_3, 3) \text{ then } d = 2 \text{ (1.0)}$$

显然这属于已有知识等价类的子集, 可以准确推导, 属于确定性规则。

3) 但是当任意给定的条件属性值  $E$  与  $F$  的比较中使得规则的可信度为零时, 则完全属于新知识, 要增加新的等价类。这也表明不能从现有规则推导实际的决策规则, 必须重新构建分量, 也就是往前面的每个向量中添加可区分的等价类, 从而才可以推导新的决策规则。对于这种新知识不但条件属性值要划分新的等价类, 决策值也要添加新的值, 这难免就存在估计的问题, 因为给定的只有条件分量的值, 没有决策属性值。因此要考虑分类质量对等价类的划分。

## 5 结束语

利用粗糙集理论在数据库中解决普遍性的问题很

难, 因为寻找约简集的算法是一个 NP 难题, 因此只能在具体问题中选择建立适当的模型来解决问题。文中创新点是: 在充分考虑区分矩阵法和基于属性重要性的启发式算法上, 用一个构造函数计算属性的加权频率, 按属性频率的大小在区分矩阵重新排序后获得约简集。通过对静态规则和决策算法的分析, 结合动态样本新增知识的各种情况, 用分类质量标识其变化的结果程度; 用集合向量包含度依次得到决策规则的极大可信度, 并形成近似决策规则。

## 参考文献:

- [1] Pawlak Z. Rough Sets[J]. International J. of Computer and Sciences, 1982, 11(5): 341-356.
- [2] 张文修, 吴伟志. 粗糙集理论与方法[M]. 北京: 科学出版社, 2001.
- [3] Zhang W X, Leung Y. Theory of including degrees and Its applications to uncertainty inferences[M] // Soft Computing in Intelligent Systems and Information Processing. New York: IEEE, 1996: 496-501.
- [4] Hu X. Knowledge discovery in databases: an attribute-oriented rough set approach[D]. Canada: University of Regina, 1995.
- [5] 胡可云. 基于概念格和粗糙集的数据挖掘方法研究[D]. 北京: 清华大学, 2001.
- [6] 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.
- [7] Kryszkiewicz M. Comparative study of alternative types of knowledge reduction in inconsistent systems[J]. International journal of Intelligence Systems, 2001, 16: 105-120.

(上接第 68 页)

## 4 结束语

提出了一种基于领域知识和数据挖掘技术的预警机制, 给出了一种预警知识描述形式, 设计并实现了预警规则挖掘算法。以现实数据为数据源, 对某高校某专业的学生成绩进行基于领域知识约束的预警规则挖掘, 获取预警规则集。以预警规则集为基础, 根据预警算法生成预警信息, 能取得较好的效果。在以后的工作中, 将研究如何根据数据的动态变化自动设置参数提高预警准确率。另外, 研究将预警项集以及项顺序关系等约束推进到频繁项集的生成过程中以提高挖掘的效率, 也是一个感兴趣的问题。

## 参考文献:

- [1] 李本建. 关于建立航空工业经济预警系统的初步设想[J]. 航空系统工程, 1993(5): 24-28.

- [2] 王耀中, 侯俊军, 刘志忠. 经济预警模型述评[J]. 湖南大学学报, 2004, 18(2): 27-31.
- [3] 柳炳祥. 基于数据挖掘的危机管理及其预警方法研究[D]. 南京: 东南大学, 2003.
- [4] 姚靠华, 蒋艳辉. 基于决策树的财务预警[J]. 系统工程, 2005, 23(10): 102-106.
- [5] 胡华平, 张 怡, 陈海涛, 等. 面向大规模网络的人侵检测与预警系统研究[J]. 国防科技大学学报, 2003, 25(1): 21-25.
- [6] Agrawal R, Imielinska T, Swami A. Mining Association Rules between Sets of Items in Large Databases[C] // Proc. of the ACM-SIGMOD 1993 Int'l Conference on Management of Data. Washington D.C.: [s.n.], 1993: 207-216.
- [7] 卢炎生, 杨 芬, 赵 栋. 带单调约束的关联规则挖掘[J]. 计算机工程, 2004, 30(15): 78-80.
- [8] 张玉林. 数据挖掘技术在教学过程和指导作用[J]. 西安通信学院学报, 2006, 5(2): 38-40.