基于粗集理论缺省数据的改进算法

王清晖, 刘文奇
(昆明理工大学 理学院,云南 昆明 650093)

摘要: 在对粗集理论研究的基础上,利用对象间的属性值差异引入粗糙相似度的概念,提出一种改进的ROUSTIDA算法,指出改进后的算法使更多的数据得到科学的补齐.实例表明此方法是比较有效的.

关键词: 粗集理论;可辨识矩阵;粗糙相似度;缺省数据
中图分类号:G202 **文献标识码:**A **文章编号:**1007-855X(2004)02-0148-03

Improved Algorithm Based on the Missing Data of Rough Set Theory

WANG Qing-hui, LIU Wen-qi
(Faculty of Science, Kunming University of Science and Technology, Kunming 650093, China)

Abstract: Based on the research on the rough set theory, an improved ROUSTIDA algorithm is proposed using the attributes difference of objects and the notion of rough set theory similarity. It is pointed out that the improved algorithm can fill much more missing data, and the results of examples exhibit its efficiency.

Key words: rough set theory; distinct matrix; rough similarity; missing data

0 引 言

1982年,Z.Pawlak提出的粗集理论[1]对数据进行分类,为处理不完整和不确定的信息提供了一种新的数学工具.ROUSTIDA算法[2]利用对象之间的无差别性对缺省数据补齐.本文以粗糙相似度作为补齐缺省数据的基础,主要讨论对决策值非空的不完备信息系统的条件属性值的补齐.

1 理论准备

定义1.1 信息表系统 $S=\langle U,A,V,f\rangle$, $A=\{a_i \mid i=1,\cdots,m\}$ 是属性集, $U=\{x_1,x_2,\cdots,x_n\}$ 是论域, $a_i(x_j)$ 是对象 x_j 在属性 a_i 上的取值, $M(i,j)$ 表示经过扩充的可辨识矩阵中第 i 行 j 列的元素,则经过扩充的可辨识矩阵 M 定义为:

$$M(i,j)=\{a_k\in A \wedge a_k(x_i)\neq a_k(x_j) \wedge a_k(x_i)\neq * a_k(x_j)\neq *\}$$

其中: $i,j=1,\cdots,n$, "$*$"表示遗失值.

定义1.2 信息表系统 $S=\langle U,A,V,f\rangle$, $A=\{a_i \mid i=1,\cdots,m\}$ 是属性集, $U=\{x_1,x_2,\cdots,x_n\}$ 是论域,设 $x_i\in U$,则对象 x_i 的遗失属性集 MAS_i,对象 x_i 的无差别对象集 NS_i 和信息系统 S 的遗失对象集 MOS 分别定义为:

$$MAS_i=\{a_k \mid a_k(x_i)=*, k=1,\cdots,m\},$$
$$NS_i=\{j \mid M(i,j)=\phi, i\neq j, j=1,\cdots,n\},$$
$$MOS=\{i \mid MAS_i\neq\phi, i=1,\cdots,n\}.$$

由于不完备信息系统中存在多个遗失值和其不同的分布,要经过多次对扩充可辨识矩阵的计算和完整化分析,对所有的遗失值进行补充,直至终止条件成立.

设初始信息系统 S^0,对象集为 $\{x_i^0\}$,相应的扩充可辨识矩阵 M^0, x_i 的遗失属性集为 MAS_i^0,无差别对象

收稿日期:2003-10-10.
第一作者简介:王清晖(1978~),女,硕士.主要研究方向:粗集理论与数据挖掘.**E-mail:**qhwang_l@yahoo.com.cn

集为 NS_i^0. 第 r 次完整化分析后的信息系统为 S^r，对象集为 $\{x_i^r\}$，相应的扩充可辨识矩阵为 M^r，x_i 的遗失属性集为 MAS_i^r，无差别对象集为 NS_i^r.

2 ROUSTIDA 算法

输入：不完备信息系统；

输出：完备的信息系统.

步骤 1：计算 M^0，MAS_i^0 和 MOS^0；令 $r = 0$.

步骤 2：

(1) 对于所有 $i \in MOS^r$，计算 NS_i^r；

(2) 产生 S^{r+1}.

1) 对于 $i \notin MOS^r$ 有 $a_k(x_i^{r+1}) = a_k(x_i^r)$，$k = 1,2,\cdots,m$；

2) 对于所有 $i \in MOS^r$，对所有 $a_k \in MAS_i^r$ 作循环：

① 如果 $card(NS_i^r) = 1$，设 $j \in NS_i^r$，若 $a_k(x_j^r) = *$，则 $a_k(x_i^{r+1}) = *$，否则 $a_k(x_i^{r+1}) = a_k(x_j^r)$；

② 否则：

(Ⅰ) 如存在 j_0 和 $j_1 \in NS_i^r$，满足 $(a_k(x_{j_0}^r) \neq *) \wedge (a_k(x_{j_1}^r) \neq *) \wedge (a_k(x_{j_1}^r) \neq a_k(x_{j_0}^r))$，则 $a_k(x_i^{r+1}) = *$；

(Ⅱ) 否则，如存在 $j_0 j_1 \in NS_i^r$，满足 $a_k(x_{j_0}^r) \neq *$，则 $a_k(x_i^{r+1}) = a_k(x_{j_0}^r)$；

(Ⅲ) 否则，$a_k(x_i^{r+1}) = *$.

(3) 如果 $S^{r+1} = S^r$，结束循环转步骤 3. 否则，计算 M^{r+1}，MAS_i^{r+1} 和 MOS^{r+1}，$r = r + 1$，转步骤 2.

步骤 3：如果信息系统还有遗失值，用其它方法处理.

步骤 4：结束.

3 对 ROUSTIDA 算法的改进

定义 3.1　信息表系统 $S = \langle U, A, V, f\rangle$，$U = \{x_1, x_2, \cdots, x_n\}$ 是论域，$A = C \cup D$ 是属性集，C 和 D 分别称为条件属性集和决策属性集，$C = \{a_i \mid i = 1,\cdots,m-1\}$，$D = \{a_m\}$. $a_j(x_j)$ 是对象 x_j 在属性 a_i 上的取值，粗糙相似度 $\rho(i,j) = \dfrac{1}{m}\sum_{k+1}^{m} card(a_k(x_i) \cap a_k(x_j))$，$0 \leq \rho \leq 1$. 其中 $i \in MOS$，$j = 1,2,\cdots,n$，$i \neq j$.　　$a_m(x_i) = a_m(x_j)$.

(规定：$a_k(x_i) = * \cup a_k(x_j) = *$，$card(a_k(x_i) \cap a_k(x_j)) = 1$，$k = 1,2,\cdots,m-1$).

此算法是比较两对象间的粗糙相似度对决策值非空的不完备信息系统的条件属性值的补齐. 确定阈值 θ，使 $\theta \leq \rho\pi 1$.

令

$$C_i = \{c \mid c = \max \rho(i,j), \theta \leq \rho\pi 1, i \in MOS, j = 1,2,\cdots,n, i \neq j\},$$

$$L_i = \{j \mid \rho(i,j) = C_i, i \in MOS, j = 1,2,\cdots,n, i \neq j\}.$$

利用粗糙相似度对 ROUSTIDA 算法的步骤 3 进行改进.

步骤 3：如果信息系统还有遗失值，则产生的不完备信息系统为 $\overline{S^0}$，计算 $\overline{MOS^0}$，$\overline{MAS_i^0}$，令 $t = 0$.

步骤 4：

(1) 对于所有 $i \in \overline{MOS^t}$，计算 $p^t(i,j)$，C_i^t 和 L_i^t；

对于所有 $i \in \overline{MOS^t}$，对于所有 $a_k \in \overline{MAS_i^t}$ 作循环：

1) 如果 $card(L_i^t) = 1$，设 $j \in L_i^t$，若 $a_k(\overline{x_j^t}) = *$，则 $a_k(\overline{x_i^{t+1}}) = *$. 否则 $a_k(\overline{x_i^{t+1}}) = a_k(\overline{x_j^t})$.

2) 否则

(Ⅰ) 如果存在 $j_0, j_1 \in L_i^t$ 满足 $(a_k(\overline{x_{j_0}^t}) \neq a_k(\overline{x_{j_1}^t})) \wedge (a_k(\overline{x_{j_0}^t}) \neq *) \wedge (a_k(\overline{x_{j_1}^t}) \neq *)$，则 $a_k(\overline{x_i^{t+1}}) = *$.

(Ⅱ) 否则,如果存在 $j_0 \in L_i^t$,满足 $a_k(\overline{x_{j_0}^t}) \neq *$,则 $a_k(\overline{x_i^{t+1}}) = a_k(\overline{x_{j_0}^t})$.

(Ⅲ) 否则,$a_k(\overline{x_i^{t+1}}) = *$.

(2) 如果$\overline{S^{t+1}} = \overline{S^t}$,结束循环转步骤 5. 否则,计算$\overline{MAS_i^{t+1}}$和$\overline{MOS^{t+1}}$,$t = t + 1$,转步骤 4.

步骤 5:结束.

4 实 例

用例子说明改进 ROUSTIDA 算法的优点,表 1 取自文献[2].

表 1 有缺省数据的信息表

U	a	b	c	d
1	1	*	1	0
2	0	1	*	0
3	1	*	2	1
4	*	1	3	1
5	0	*	*	1

表 2 用 ROUSTIDA 算法(主要指前两步)补齐信息表

U	a	b	c	d
1	1	*	1	0
2	0	1	*	0
3	1	*	2	1
4	0	1	3	1
5	0	1	3	1

表 3 用改进后的算法补齐信息表($\theta = 0.5$)

U	a	b	c	d
1	1	1	1	0
2	0	1	1	0
3	1	1	2	1
4	0	1	3	1
5	0	1	3	1

通过各表比较分析可知,ROUSTIDA 算法大大降低了运算复杂度,同时便于处理大批量缺省数据,缺点是应用范围不够,对于许多情形无能为力.改进的 ROUSTIDA 算法,使更多的缺省数据得到科学补齐.

5 结 论

本文提出一种基于粗集理论,利用对象间的粗糙相似度,对 ROUSTIDA 算法改进的一种缺省数据的补齐方法,此方法使更多的数据得到科学的补齐.不同的问题,要求选择适当的阈值.但是,本文没有涉及该算法的修正,还需作进一步的研究.

参考文献:

[1] PAWLAKZ. Rough sets[J]. Int J of Computer and Information Science, 1982, 11(5): 341 ~ 356.

[2] 王国胤. Rough 集理论与知识获取[M]. 西安:西安交通大学出版社, 2001.

[3] 张文修,吴文志,梁吉业,等. 粗糙集理论与方法[M]. 北京:科学出版社, 2001.

[4] 石峰,娄臻亮,张永清,等. 基于模糊 - 粗糙集模型的一种归纳学习方法[J]. 上海交通大学学报, 2002, 36(7): 920 ~ 924.