多决策表缺失属性补齐算法的研究

焦 娜 苗夺谦 张红云
(同济大学嵌入式系统与服务计算教育部重点实验室 上海 201804)
(同济大学电子与信息工程学院计算机科学与技术系 上海 201804)
(国家高性能计算机工程中心同济分中心 上海 201804)

摘 要 不完备数据是造成信息系统不确定的主要原因之一,对数据挖掘、知识发现等造成了困难。已有的大多数不完备数据的填补算法主要考虑单个决策表的情况,有关多决策表缺失属性补齐算法却报道不多。为此,首先定义了多决策表的属性综合重要性;并以此为启发式信息,基于多决策表的内在关联性,依次补齐缺失属性;最后,实验证明该算法是有效可行的。

关键词 粗糙集理论,多决策表,缺失属性,补齐算法

Research on Algorithm for Completing Missing Attributes in Multiple Decision Tables

JIAO Na MIAO Duo-qian ZHANG Hong-yun
(Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 201804, China)
(Department of Computer Science and Technology, Tongji University, Shanghai 201804, China)
(Tongji Branch, National Engineering & Technology Center of High Performance Computer, Shanghai 201804, China)

Abstract Data mining and knowledge discovery on incomplete data is difficult but inevitable and uncertain. The previous methods of dealing with incomplete data were developed under a single decision table, and not under multiple decision tables. The general significance of attributes in multiple decision tables is defined; according to the heuristic information, an algorithm for completing missing attributes in multiple decision tables was proposed. Missing attributes and corresponding values calculated were added into decision tables including missing attributes. Preliminary data from the experiments show that the algorithm is effective to analyze multiple decision tables.

Keywords Rough set theory, Multiple decision tables, Missing attributes, Completing algorithm

1 引言

波兰数学家 Pawlak 在 20 世纪 80 年代初提出粗糙集理论[1,2],由于它能有效地分析和处理不精确、不一致、不完整等各种不完备信息,该理论在机器学习、数据挖掘及模式识别等多个领域得到了广泛的应用[3]。近年来,利用粗糙集理论对不完备信息的研究逐渐受到研究人员的重视,并提出了几种处理不完备信息的方法[4,5]。目前,填补不完备数据的方法有均值法、最大频率法[4]、断点法[6]等,但这些方法都是针对单个信息系统或单个决策表,有关多信息系统或多决策表的情况却报道不多。

数据库越来越庞大,信息系统和决策表也越来越多,而且很多信息系统和决策表是有一定关联的。在单个信息系统或单个决策表基础上粗糙集理论的研究不再适合于多信息系统或多决策表。为此,文献[7-9]提出了几种对于多决策表的研究方法。文献[7]提出了在多决策表中,属性相同,对象有所不同,根据多决策表的内在关联性补齐了多决策表的算法。但该算法并未提及在多决策表中存在缺失属性的情况。

在多决策表中存在缺失属性的情况会经常出现。当信息来自不同的信息源,很可能得到多个决策表,或者当每个决策表中的对象和属性是由不同的多个决策者做出的时候,很可能有决策者的个人偏好在里面,也有可能得到多个决策表。而这些决策表里面既包含相同的属性也包含不同的属性,因此做出的决策表很可能在个别的属性上有所缺失,这对数据挖掘造成了很大的困难。本文首先定义了多决策表的属性综合重要性,并以此为启发式信息,提出多决策表缺失属性补齐算法。该算法根据多决策表的内在关联性,依次补齐缺失属性。最后,实验证明该算法是有效可行的。

2 基本概念

一般情况下,一个决策表可以定义为四元组$\langle U, C\cup\{d\}, V, f\rangle$,其中$U=\{x_1,\cdots,x_n\}$为论域;$C,d$分别为条件属性集和决策属性;$V=\bigcup_{a\in C\cup\{d\}} V_a$,$V_a$表示属性$a$的值域;$f: U\times(C\cup\{d\})\to V$是一个信息函数,即对$\forall x\in U, a\in C\cup\{d\}$,有$f(x,a)\in V_a$。便于计算,本文用到了文献[7]给出的决策表表示方法。得到如下定义:

到稿日期:2008-02-22 本文受国家自然科学基金(60475019,60775036),教育部博士学科点专项科研基金(20060247039)资助。
焦 娜(1977-),女,博士研究生,研究方向为模式识别与智能系统、粗糙集,E-mail:zdx.jn@163.com;苗夺谦(1964-),男,教授、博士生导师,研究方向为人工智能、模式识别、粗糙集;张红云(1972-),女,讲师,研究方向为模式识别与智能系统。

定义 1 设 d 为决策表的决策属性，V_d 表示决策属性 d 的值域，x 表示对象，则上面的决策表可用四元组 $\langle W, C\cup\Sigma, V, f\rangle$ 重新表示，其中 $W=\{w_1,\cdots,w_t\}$ 为模式集合，$w_i=\bigcup_{a\in C}\{\langle a,a(x)\rangle\}$ 为模式，$a(x)$ 表示对象 $x\in U$ 在属性 $a\in C$ 上的值；C 为条件属性集；$\Sigma=\{\sigma_C(w_i,v_d),\forall v_d\in V_d\}$ 表示模式出现频率集，其中 σ_C 表示频率函数，$\sigma_C(w_i,v_d)$ 表示模式为 w_i、决策属性值为 v_d 的出现频率；$V=\bigcup_{a\in C}V_a$，V_a 表示属性 a 的值域；$f:W\times C\to V$ 是一个信息函数，即对 $\forall w\in W, a\in C$，有 $f(w,a)\in V_a$。

表 1 是一个普通决策表。根据定义 1，表 1 中的决策表重新表示为表 2。表 2 中每个模式出现频率集就可以用向量 $(\sigma_C(w_i,accept),\sigma_C(w_i,reject))$ 来表示。对粗糙集理论的分析中，对象在决策表中出现的顺序并不影响分析的结论，因此表 2 和表 1 是等价的。本文的决策表都是基于模式来讨论的。

表 1 决策表

U	Design	Function	Size	Dec
x_{11}	Classic	Simple	Compact	Accept
x_{12}	Classic	Simple	Compact	Accept
x_{13}	Classic	Simple	Compact	Reject
x_{21}	Classic	Multiple	Normal	Accept
x_{22}	Classic	Multiple	Normal	Accept

表 2 重新表示的决策表

W	Design	Function	Size	Σ
w_1	Classic	Simple	Compact	(2,1)
w_2	Classic	Multiple	Normal	(2,0)

定义 2 根据定义 1，一个决策表集合可以定义如下：

$\Gamma=\{T_i, i=1,\cdots,h\}$

其中 $T_i=\langle W_i, C_i\cup\Sigma_i, V_i, f_i\rangle$，$C_i$ 是 T_i 的条件属性；

$C(\Gamma)=\bigcup_{i=1}^{h}C_i$ 表示 Γ 中出现的所有属性的集合；

$A_i=\{a\mid a\notin C_i\wedge a\in C\}$ 表示 T_i 的缺失属性集合；

$A(\Gamma)=\bigcup_{i=1}^{h}A_i$ 是 Γ 中所有的缺失属性集合；

$Z(\Gamma,a)=\{T_i\mid a\in C_i, \forall T_i\in\Gamma\}$ 表示 Γ 所有包含属性 a 的决策表的集合；

$Y(\Gamma,a)=\{T_i\mid a\notin C_i, \forall T_i\in\Gamma\}$ 表示 Γ 所有缺失属性 a 的决策表的集合。

定义 3 设 w_i 是决策表 T_i 中的模式，W_j 是决策表 T_j 的模式集合，$i\neq j$，则定义 T_j 中与 w_i 的重合条件属性值集合为

$B(w_i,W_j)=\{\langle a,a(w_i)\rangle\mid\forall a\in C_i\cap C_j, \exists w\in W_j, a(w_i)=a(w)\}$

定义 T_j 中能够与 $B(w_i,W_j)$ 完全匹配的模式集合为 $F(w_i,W_j)=$

$\{w\mid\forall\langle a,v_a\rangle\in B(w_i,W_j), \forall w\in W_j, v_a=a(w)\}$

定义 $F(w_i,W_j)$ 中能够匹配 $\langle b,v_b\rangle$ 的模式集合

$L(\langle b,v_b\rangle,w_i,W_j)=\{w\mid\forall w\in F(w_i,W_j), b(w)=v_b\}$。

3 多决策表缺失属性补齐算法

在多个决策表中，各个决策表的缺失属性可能有所不同。例如，一个决策表可能缺失多个属性，或者几个决策表都缺失同一个属性，这需要依次补齐缺失属性。但是缺失属性的重要性又有所不同，先添加重要性低的缺失属性，可能会对后面添加重要性高的属性有所影响。因此需要对缺失属性进行排序，依次添加较重要属性。

3.1 多决策表中的属性综合重要性度量

实际的数据集往往都受到一定噪声的干扰。为了正确地处理这些噪声，本文引入可变精度粗糙集模型。

定义 4[10,11] 对于 $T_i\in\Gamma$，则对于任意属性 $a\in C_i$ 的重要性 $SGF(a,W_i)$ 定义为

$$SGF(a,W_i)=\frac{|POS^{\beta}_{C_i}(\Sigma_i)|-|POS^{\beta}_{C_i\setminus\{a\}}(\Sigma_i)|}{|W_i|}$$

其中 $|POS^{\beta}_{C_i}(\Sigma_i)|=\sum_{\forall w\in W_i,\exists v_d\in V_d,\sigma_{C_i}(w,v_d)/\sum_{\forall v_d\in V_d}\sigma_{C_i}(w,v_d)\geqslant 1-\beta}\sigma_{C_i}(w,v_d)$，$\beta\in[0,0.5)$ 是可变精度粗糙集中依赖于噪声的一个系数。

定义 5 在 Γ 中，属性 a 的综合重要性 $S(a)$ 定义如下：

$$S(a)=\frac{\sum_{i=1}^{h}|W_i|SGF(a,W_i)}{\sum_{i=1}^{h}|W_i|}$$

$S(a)$ 的值越大，说明属性 a 就越重要。本文将 $S(a)$ 作为启发式信息，可以减少搜索空间。

3.2 多决策表缺失属性补齐算法

下面介绍如何将属性 a_m 同时添加到所有缺失该属性的决策表 $Y(\Gamma,a_m)$ 中去。

首先，添加了属性 a_m 后，新决策表中的模式将有所改变。假设 a_m 的值域为 $V_{a_m}=\{v^1_{a_m},\cdots,v^r_{a_m}\}$。添加了属性 a_m 后的每个模式都变成 r 个模式。例如，原来的一个模式 w_u 添加 a_m 后变成

$$w_{u1}=\{w_u(c_1),w_u(c_2),\cdots,w_u(c_q),v^1_{a_m}\}$$
$$w_{u2}=\{w_u(c_1),w_u(c_2),\cdots,w_u(c_q),v^2_{a_m}\}$$
$$\vdots$$
$$w_{ur}=\{w_u(c_1),w_u(c_2),\cdots,w_u(c_q),v^r_{a_m}\}$$

其中 $\{w_u(c_1),w_u(c_2),\cdots,w_u(c_q)\}$ 为原来 w_u 对应的各个属性值。

其次，一个模式 w_u 分散成多个模式 $w_{u1},w_{u2},\cdots,w_{ur}$，为了保持原来数据的正确，$w_u$ 的频率 σ_{w_u} 也应分配到 r 个模式中去。$w_{u1},w_{u2},\cdots,w_{ur}$ 的频率需要根据决策表之间的内在关联性来计算。

对于任意的 w_u 和 $W_j\in Z(\Gamma,a_m)$，根据定义 3 求出 T_j 中与 w_u 的重合条件属性值集合 $B(w_u,W_j)$ 及其模式集合 $F(w_u,W_j)$，然后求出 $F(w_u,W_j)$ 只能够匹配 $\langle a_m,v^i_{a_m}\rangle$ 的模式集合 $L(\langle a_m,v^i_{a_m}\rangle,w_u,W_j)$，以 $L(\langle a_m,v^i_{a_m}\rangle,w_u,W_j)$ 在 $F(w_u,W_j)$ 中的比例作为权值把原模式 w_u 的频率分配到新模式中去，分配权值可如下定义：

$$\mu(w_u,v_d,W_j,\langle a_m,v^i_{a_m}\rangle)=\begin{cases}\sum_{w\in L}\sigma(w,v_d)/\sum_{w\in F}\sigma(w,v_d), & \text{if } F\neq\emptyset\\ 0, & \text{if } F=\emptyset\end{cases}\quad(1)$$

其中 $F=F(w_u,W_j)$，$L=L(\langle a_m,v^i_{a_m}\rangle,w_u,W_j)$。

如果 Γ 中有多个决策表含有属性 a_m，则用相同的方法求得对应模式在每个含有属性 a_m 的决策表中 r 个值的分配比例，最后用分配比例的均值与原来一个模式的出现频率相乘，即为新表的 r 个模式的频率。计算公式见(2)。令 $i=1,\cdots,r$，

$$\sigma(w_{ui},v_d)=\sigma(w_u,v_d)\times\frac{\sum_{j=1}^{|Z(\Gamma,a_m)|}\mu(w_u,v_d,W_j,\langle a_m,v^i_{a_m}\rangle)}{|Z(\Gamma,a_m)|}\quad(2)$$

这里只将 w_{ia} 的频率按照其权值 $\mu(w_u, v_d, W_j, \langle a_m, v_{a_m}^i \rangle)$ 将其频率 $\sigma(w_u, v_d)$ 分配到新模式中去，并没有改变原有信息，在保持了原表中信息完整性的基础上添加了缺失属性。

最后，整理新决策表，删除模式频率都是 0 的模式，得到最终添加完属性 a_m 的新决策表。依此类推，使用同样的方法补齐其他有缺失属性的决策表。

算法 1 多决策表缺失属性补齐算法

输入：不完备决策表集合 $\Gamma=\{T_i, i=1,\cdots,h\}$，

其中 $T_i=\langle W_i, C_i \cup \sum_i, V_i, f_i \rangle$；$C=\bigcup_{i=1}^{h} C_i$；

$A_i=\{a \mid a \notin C_i \wedge a \in C\}$ 表示 T_i 的缺失属性；

$A(\Gamma)=\bigcup_{i=1}^{h} A_i$ 是 Γ 中所有的缺失属性集合；给定阈值 β $(0 \leqslant \beta < 0.5)$。

输出：完备的决策表集合 $\Gamma'=\{T_i', i=1,\cdots,h\}$，其中 $T_i'=\langle W_i', C \cup \sum_i', V_i', f_i' \rangle$。

步骤 1　令 $\Gamma'=\Gamma, C'=C, A'=A$。

步骤 2　如果 $A'=\emptyset$，转向步骤 5，否则转向步骤 3。

步骤 3　根据定义 5 得到当前决策表集合 Γ' 的最重要缺失属性 $a_m=\underset{a \in A'}{\operatorname{argmax}}\{S(a)\}$。

步骤 4　根据前面所讲的算法将 a_m 添加到所有缺失该属性的决策表 $Y(\Gamma, a_m)$ 中。

①首先，确定在所有缺失属性 a_m 的决策表中添加 a_m 后的新模式。

②然后，计算这些新的模式在所有决策属性值上的出现频率。

③整理新决策表，删除模式出现频率都是 0 的模式。

④令 $C'=C' \cup \{a_m\}, A'=A' \setminus \{a_m\}$，转向步骤 2。

步骤 5　Γ' 即为所求，结束。

4　实例分析

4.1　实例

根据上述算法，用 UCI 的机器学习数据库中 car 数据库作为样本数据库[12] 来进行验证。选取 {Buying, Maint, Doors, Persons, Lug-boot, Safety, Dec} 6 个条件属性和 1 个决策属性及部分对象构造了 4 个决策表，分别用 $a_1, a_2, a_3, a_4, a_5, a_6$ 表示决策表的属性名，用 $V_{a_1}=V_{a_2}=V_{a_6}=\{h, l\}$ 表示 a_1, a_2, a_6 的属性值{high, low}，用 $V_{a_4}=\{b, s\}$ 表示 a_4 的属性值{*big*, *small*}，d=Dec. (Decision)，$V_d=\{acc, unacc\}$。每个决策表的 Dec. 用模式出现频率的向量来表示，$\Sigma=(\sigma_C(w_i, acc), \sigma_C(w_i, unacc))$。$\Gamma=\{T_1, T_2, T_3, T_4\}$，见表 3。假设 4 个决策表是由 4 个决策者对汽车做出的评估，这些决策者对于汽车的评估都有各自的偏好。这些有缺失属性的不完备决策表集是上述完备决策表集随机丢失 4 个属性后形成的，其中 T_1, T_2 缺失属性 a_1，T_3 缺失属性 a_3，T_4 缺失属性 a_5，见表 4。

$C_1=\{a_2, a_3, a_4, a_5, a_6\}, C_2=\{a_2, a_3, a_4, a_5, a_6\}, C_3=\{a_1, a_2, a_4, a_5, a_6\}, C_4=\{a_1, a_2, a_3, a_4, a_6\}$。$C(\Gamma)=\{a_1, a_2, a_3, a_4, a_5, a_6\}$；$A_1=\{a_1\}, A_2=\{a_1\}, A_3=\{a_3\}, A_4=\{a_5\}$。$A(\Gamma)=\{a_1, a_3, a_5\}$；$Z(\Gamma, a_1)=\{T_3, T_4\}, Z(\Gamma, a_2)=\Gamma, Z(\Gamma, a_3)=\{T_1, T_2, T_4\}, Z(\Gamma, a_4)=\Gamma, Z(\Gamma, a_5)=\{T_1, T_2, T_3\}, Z(\Gamma, a_6)=\Gamma; Y(\Gamma, a_1)=\{T_1, T_2\}, Y(\Gamma, a_2)=\emptyset, Y(\Gamma, a_3)=\{T_3\}, Y(\Gamma, a_4)=\emptyset, Y(\Gamma, a_5)=\{T_4\}, Y(\Gamma, a_6)=\emptyset$。

表 3　原多决策表集合

表 3.1

w_1	a_1	a_2	a_3	a_4	a_5	a_6	Σ
w_1	h	h	2	4	b	h	(3,25)
w_2	l	h	2	4	b	h	(24,8)
w_3	h	h	2	4	b	l	(8,12)
w_4	l	h	2	4	b	l	(12,19)
w_5	l	h	2	2	b	h	(10,19)
w_6	l	l	2	4	s	h	(30,0)
w_7	l	l	4	4	b	h	(24,0)

表 3.2

w_2	a_1	a_2	a_3	a_4	a_5	a_6	Σ
w_1	h	h	2	4	b	h	(3,24)
w_2	l	h	2	4	b	h	(20,7)
w_3	h	h	2	4	b	l	(9,15)
w_4	l	h	2	4	b	l	(13,22)
w_5	l	h	2	2	b	h	(8,18)
w_6	l	l	2	4	s	h	(20,1)
w_7	l	l	4	4	b	h	(25,0)

表 3.3

w_3	a_1	a_2	a_3	a_4	a_5	a_6	Σ
w_1	h	h	2	4	b	h	(3,24)
w_2	l	h	2	4	b	h	(18,8)
w_3	h	h	2	4	b	l	(9,19)
w_4	l	h	2	4	b	l	(10,23)
w_5	l	h	2	2	b	h	(6,16)
w_6	l	l	2	4	s	h	(15,0)
w_7	l	l	4	4	b	h	(17,0)

表 3.4

w_4	a_1	a_2	a_3	a_4	a_5	a_6	Σ
w_1	h	h	2	4	b	h	(2,23)
w_2	l	h	2	4	b	h	(16,6)
w_3	h	h	2	4	b	l	(10,18)
w_4	l	h	2	4	b	l	(14,24)
w_5	l	h	2	2	b	h	(10,20)
w_6	l	l	2	4	s	h	(21,1)
w_7	l	l	4	4	b	h	(26,0)

选 $\beta=0.2$，根据定义 5，决策表 T_1, T_2 的缺失属性 a_1 的综合重要性是 0.131，决策表 T_3 的缺失属性 a_3 的综合重要性是 0，决策表 T_4 的缺失属性 a_5 的综合重要性是 0。得出最重要缺失属性是 a_1，把属性 a_1 同时添加到决策表 T_1, T_2 中。属性 a_1 的值域为 $V_{a_1}=\{h, l\}$，那么添加了属性 a_1 后的每个模式就对应变成了 2 个模式，见表 5。

接下来确定各个模式的频率，对于 T_1 的模式 w_{11} 按照上面定义 3，得到 $B(w_1, W_3)=\{\langle a_2, h\rangle, \langle a_4, 4\rangle, \langle a_5, b\rangle, \langle a_6, h\rangle\}$，$F(w_1, W_3)=\{e_1, e_3\}$，$L(\langle a_1, h\rangle, w_1, W_3)=\{e_1\}$，$L(\langle a_1, l\rangle, w_1, W_3)=\{e_3\}$。

计算 $\mu(w_1, acc, W_3, \langle a_1, h\rangle)=\dfrac{\sigma(e_1, acc)}{\sigma(e_1, acc)+\sigma(e_3, acc)}=\dfrac{3}{3+18}=0.143$，用同样方法求得 $\mu(w_1, acc, W_4, \langle a_1, h\rangle)=\dfrac{2}{2+16}=0.111$，最后得到 T_1 的模式 w_{11} 在决策属性值为 acc 的出现频率 $\sigma_{C'_1}(w_{11}, acc)=27*(0.143+0.111)/2=3.43$，相应得 $\sigma_{C'_1}(w_{11}, unacc)=25.46$，这样决策表 T_1 的模式 w_{11} 在所有决策属性值上的出现频率为(3.43, 25.46)，并删除频率为 0 的模式。

同理，添加其他的缺失属性，得到的最终结果如表 6 所示。

表 4　随机去掉 4 个属性后多决策表

表 4.1　决策表 T_1

w_1	a_2	a_3	a_4	a_5	a_6	Σ
w_1	h	2	4	b	h	(27,33)
w_2	h	2	4	b	l	(20,31)
w_3	h	2	2	b	h	(10,19)
w_4	l	2	4	s	h	(30,0)
w_5	l	4	4	b	h	(24,0)

表 4.2　决策表 T_2

w_2	a_2	a_3	a_4	a_5	a_6	Σ
w_1	h	2	4	b	h	(23,31)
w_2	h	2	4	b	l	(22,37)
w_3	h	2	2	b	h	(8,18)
w_4	l	2	4	s	h	(20,1)
w_5	l	4	4	b	h	(25,0)

表 4.3　决策表 T_3

w_3	a_1	a_2	a_4	a_5	a_6	Σ
e_1	h	h	4	b	h	(3,24)
e_2	h	h	4	b	l	(9,19)
e_3	l	h	4	b	h	(18,8)
e_4	l	h	4	b	l	(10,23)
e_5	l	h	2	b	h	(6,16)
e_6	l	l	4	s	h	(15,0)
e_7	l	l	4	b	h	(17,0)

表 4.4　决策表 T_4

w_4	a_1	a_2	a_3	a_4	a_6	Σ
f_1	h	h	2	4	h	(2,23)
f_2	h	h	2	4	l	(10,18)
f_3	l	h	2	4	h	(16,6)
f_4	l	h	2	4	l	(14,24)
f_5	l	h	2	2	h	(10,20)
f_6	l	l	2	4	h	(21,1)
f_7	l	l	4	4	h	(26,0)

对补齐缺失属性的4个表的属性值取整。将填补后的决策表与原表进行对比，T_1'表的错误率为 3/194 * 100%= 1.55%，T_2'表的错误率为 2/185 * 100%= 1.08%，T_3'表的错误率为 6/168 * 100%=3.57%，T_4'表的错误率为 2/191= 1.05%。平均错误率 13/738=1.76%，其填补效果比较满意。通过示例验证了利用上述算法对不完备多决策表进行处理的可行性。

表5 T_1 添加属性 a_1 后 w_1 变成两个模式

	a_2	a_3	a_4	a_5	a_6	a_1	Σ
w_{11}	h	2	4	b	h	h	(3.43,25.46)
w_{12}	h	2	4	b	h	l	(23.57,7.54)
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

表6 补齐后的决策表集合

表6.1 补齐属性 a_1 并整理后的 T_1'

W_1'	a_1	a_2	a_3	a_4	a_5	a_6	Σ
g_1	h	h	2	4	b	h	(3.43,25.46)
g_2	l	h	2	4	b	h	(23.57,7.54)
g_3	h	h	2	4	b	l	(8.9,13.65)
g_4	l	h	2	4	b	l	(11.1,17.35)
g_5	l	h	2	2	b	h	(10,19)
g_6	l	l	2	4	s	h	(30,0)
g_7	l	L	4	4	b	h	(24,0)

表6.2 补齐属性 a_1 并整理后的 T_2'

W_2'	a_1	a_2	a_3	a_4	a_5	a_6	Σ
g_1	h	h	2	4	b	h	(2.92,23.92)
g_2	l	h	2	4	b	h	(20.08,7.08)
g_3	h	h	2	4	b	l	(9.79,16.3)
g_4	l	h	2	4	b	l	(12.21,20.7)
g_5	l	h	2	2	b	h	(8,18)
g_6	l	l	2	4	s	h	(20,1)
g_7	l	l	4	4	b	h	(25,0)

表6.3 补齐属性 a_3 并整理后的 T_3'

W_3'	a_1	a_2	a_3	a_4	a_5	a_6	Σ
g_1	h	h	2	4	b	h	(3,24)
g_2	l	h	2	4	b	h	(18,8)
g_3	h	h	2	4	b	l	(9,19)
g_4	l	h	2	4	b	l	(10,23)
g_5	l	h	2	2	b	h	(6,16)
g_6	l	l	2	4	s	h	(12.23,0)
g_7	l	l	4	4	b	h	(14.47,0)
g_8	l	l	2	4	b	h	(2.53,0)
g_9	l	l	4	4	s	h	(2.77,0)

表6.4 补齐属性 a_5 并整理后的 T_4'

W_4'	a_1	a_2	a_3	a_4	a_5	a_6	Σ
g_1	h	h	2	4	b	h	(2,23)
g_2	l	h	2	4	b	h	(16,6)
g_3	h	h	2	4	b	l	(10,18)
g_4	l	h	2	4	b	l	(14,24)
g_5	l	h	2	2	b	h	(10,20)
g_6	l	l	2	4	s	h	(19.8,1)
g_7	l	l	4	4	b	h	(24.61,0)
g_8	l	l	2	4	b	h	(1.2,0)
g_9	l	l	4	4	s	h	(1.39,0)

4.2 实验及分析

用UCI的car数据库作为样本数据库，选取属性与上面的实例相同。选取1600个对象构造成4个决策表，分别随机去掉1,2,3,4,5,6,7,8个属性。用本算法补齐，得到的结果如图1所示。

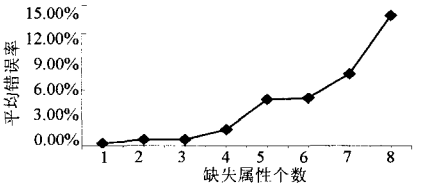

图1 缺失属性个数与平均错误率的对比关系

通过实验可以发现，随着缺失属性个数的增多，补齐后的平均错误率的总体趋势也会越高。如果缺失属性过多，即使补齐了所有缺失的属性，也会有很大误差，这样就没有意义了。因此，各个决策表中的缺失属性不能太多。

结束语 本文在保持原决策表信息不变的情况下，利用决策表之间信息的关联性，补齐多决策表缺失属性。即保持了原决策表集合信息的完整性，又使决策表集合中的所有决策表属性一致，使数据更能适用于现有的数据挖掘方法，为不完备多决策表数据挖掘的预处理提供了一条新的思路。

参考文献

[1] Pawlak Z. Rough Sets[J]. International Journal of Information Computer Science, 1982, 11(5): 341-356

[2] Pawlak Z, Grzymala-Busse J, Slowinski R, et al. Rough sets. Communications of ACM, 1995, 38(11): 89-95

[3] Greco S, Inuiguchi M, Slowinski R. Fuzzy rough sets and multiple-premise gradual decision rules[J]. International Journal of Approximate Reasoning, 2006, 41(2): 179-211

[4] Kryszkiewicz M. Rough set approach to incomplete information system[J]. Information Sciences, 1998, 112: 39-49

[5] Kryszkiewicz M. Rules in incomplete information systems[J]. Information Sciences, 1999, 113: 271-292

[6] 鄂旭，高学东，武森，等. 信息表中不完备数据的填补方法[J]. 北京科技大学学报，2005，27(3)：364-366

[7] Inuiguchi M, Suzuki J, Miyajima T. Variable Precision Rough Set Approach to Multiple Decision Tables[C]// Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing (RSFDGrC). Regina, Canada, Aug. 31 - Sep. 3, 2005: 304-313

[8] Inuiguchi M, Suzuki J, Miyajima T. Toward Rule Extraction from Multiple Decision Tables Based on Rough Set Theory[C]// Proceedings of 15th Mini-Euro Conference on Managing Uncertaining in Decision Support Models (2004) CD-ROM. Coimbra, Portuga l, Sep. 2004

[9] Milton R S, Maheswari V U, Siromoney A. Studies on Rough Sets in Multiple Tables[C]// Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing (RSFDGrC). Regina, Canada, Aug. 31 - Sep. 3, 2005: 265-274

[10] 王国胤. Rough集理论与知识获取[M]. 西安：西安交通大学出版社，2001：51-52

[11] Ziarko W. Variable Precision Rough Set Model[J]. Journal of Computer and System Sciences, 1993, 46: 39-59

[12] UCI. 机器学习数据库. ftp://ftp.ics.uci.edu/pub/machine-learning-databases