第 22 卷 第 3 期
2007 年 9 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol.22 No. 3
Sep. 2007

一种不完备决策表的数据补齐方法

王希雷
（天津科技大学计算机科学与信息工程学院，天津 300222）

摘 要：针对不完备信息系统，提出缺失比概念. 用缺失比表示待填充样本中缺失数据和剩余数据对决策结果的影响能力的大小，进而根据缺失比选择使用的约简，然后通过填充矩阵寻找缺失数据的最大可能值. 采用对 Rough 集进行扩充和数据填补相结合的决策表数据补齐方法. 试验结果表明该方法有较好效果.

关键词：Rough 集；数据补齐；不完备信息系统；遗漏值
中图分类号：X948 **文献标识码**：A **文章编号**：1672-6510（2007）03-0062-03

A Data Complement Method for Incomplete Decision Tables

WANG Xi-lei
（College of Computer Science and Information Engineering, Tianjin University of Science & Technology, Tianjin 300222, China）

Abstract: The concept of miss-ratio was defined. Miss-ratio is a ratio about data between missing and leaving. It can be used to express the effect for decisions in incomplete information systems. And we can choose the reduction based on miss-ratio and search optimization values for missing data based on filling-matrix. The data complement method for incomplete decision tables was proposed based on rough set extension and data complement. Experiment results show that this method is effective.

Keywords: rough sets; data complementation; incomplete information systems; missing values

Pawlak 教授提出的经典 Rough 集理论在知识发现领域取得了很大的成功. 但它是基于完备信息系统的假设，而现实数据集大都是不完备信息系统. 为使 Rough 集能适应不完备信息系统，目前主要有两类方法，一是把不完备信息系统转化为完备信息系统，即数据补齐[1]. 目前数据补齐方法是，直接删除含缺失数据的样本，通过统计等方法对缺失数据进行估计补充，用已有数据作为训练样本建立预测模型预测缺失数据[2]；二是对经典 Rough 集理论中的相关概念在不完备信息系统下进行适当扩充[3]. 该方法的基本思想是用一个特殊值代替缺失值，假定该特殊值具有某种性质，例如假定该特殊值可以是任何可能的值. 通过统计方法对缺失数据进行补充的数据补齐方法具有很快的运行速度，但是填补数据质量很差. 目前建立模型预测缺失数据方法中的不足之处是对于缺失数据过多的样本无法得到满意结果，而且建模过程使用传统的规则提取方法，规则质量与基于 Rough 集方法得到的规则比较，质量不高. 对经典 Rough 集扩充的方法在处理缺失数据较多样本时也存在同样问题，而且整体结果也不令人满意. 总体上说，由于 Rough 集理论及应用的研究成为热点的时间不是很长，目前研究者大都把精力集中在特征提取等核心问题的研究，而对于属于预处理的不完备数据处理的研究较少，所以目前不完备信息处理的质量不令人满意. 而预处理的质量直接影响特征提取的结果，是基于 Rough 集知识发现的一个重要环节，随着对特征提取等核心问题研究的深入，预处理的研究必将成为 Rough 集理论及应用领域的下一个研究热点.

本文提出一种基于 Rough 集二进制可辨矩阵的不完备决策表数据补齐方法，该方法结合使用对 Rough 集扩充和数据补齐 2 种方法，针对缺失数据较多的情况，提出缺失比的定义，给出了缺失数据对决策结果影响的数值表示形式，为寻找缺失数据对决策结果影响最小的模型提供了依据. 定义了填充矩阵，

收稿日期：2006-12-19；**修回日期**：2007-05-22
基金项目：天津科技大学自然科学基金资助项目(20050226)；天津高等学校科技发展计划基金资助项目(04310951R)
作者简介：王希雷（1972—），男，黑龙江人，讲师，硕士.

用以准确寻找缺失数据最大可能的值.

1　二进制可辨矩阵

定义 1：二进制可辨矩阵的行定义为对象对，表示论域 U 中 2 个对象 u_i 与 u_j 的对应属性值比较结果[4].

决策表是一种常用的信息系统，它表示当满足某些条件时，决策行为应当如何进行. 设决策表为 $T=(U,C,D)$，其中论域 $U=\{u_1,u_2,\cdots,u_n\}$，条件属性集 $C=\{c_1,c_2,\cdots,c_m\}$，决策属性 $D=\{d\}$，则决策表 T 对应的二进制可辨矩阵 $\boldsymbol{M}_t$ 构造如下：

矩阵的每一列对应一个条件属性，共有 m 列，每一行对应一个对象对（u_i,u_j），其中 u_i 和 u_j 的决策属性 $d(u_i)\neq d(u_j)$，即这一对对象属于不同的决策类. 设 T 对应的二进制可辨矩阵为 $\boldsymbol{M}_t$，则 $\boldsymbol{M}_t$ 的行对应在不可分辨关系 IND(A)下的可分辨的对象对（u_i,u_j），它的列对应属性集 C 中的属性 c_{ij}，设 $\boldsymbol{M}_t=(m_{(i,j),q})$

$$m_{(i,j),q}=\begin{cases}1 & c_q(u_i)\neq c_q(u_j)\\ 0 & c_q(u_i)=c_q(u_j)\end{cases}$$

定义 2：决策表 $T(U,C,D)$对应的二进制可辨矩阵 $\boldsymbol{M}_t$ 中某列 c_i 对应的 1 的个数定义为属性 c_i 的区分能力，表示 c_i 能区分开的对象对的数量；集合 $C'\subseteq C$对应列逻辑加后 1 的个数定义为属性集 C'的区分能力，表示 C'能区分开的对象对的数量.

例 1. 设 $C'=\{c_1,c_2\}$，$c_1=(1,1,0,1)^{\mathrm{T}}$，$c_2=(0,1,0,1)^{\mathrm{T}}$，$c_1+c_2=(1,1,0,1)^{\mathrm{T}}$，则 C'的区分能力为 3.

定义 3：决策表 $T(U,C,D)=\{T_1,T_2\}$，其中 $T_1=\{u_i\mid u_i\in T$ 且 u_i 不含缺失属性值$\}$，$T_2=\{u_i\mid u_i\in T$ 且 u_i 含缺失属性值$\}$. 设 T_1 的一个约简为 R，T_2 中某个样本 u_i 中缺失属性值对应属性的集合为 S，$S\cap R=R$，则 R'的区分能力与$\{R-R'\}$的区分能力的比值定义为 u_i-R 的缺失比. 表示样本 u_i 的缺失值在 R 中对应属性的区分能力与 R 中剩余属性的区分能力的比值.

通过缺失比的定义，可以得到缺失属性对决策结果影响的数值表达形式.

定义 4：对决策表 $T=(U,C,D)$,其中论域 $U=\{u_1,u_2,\cdots,u_n\}$，条件属性集 $C=\{c_1,c_2,\cdots,c_m\}$，决策属性 $D=\{d\}$，其中缺失值填 null. 设 u'为待填充样本，a_j 为 u'中待填充属性值对应属性. 构造一个矩阵 $\boldsymbol{M}_r$，设 $\boldsymbol{M}_r$ 中行对应 T 中一个样本 u_i 与待填充样本 u'的比较结果，$\boldsymbol{M}_r$ 的前 m 列对应 c_1 到 c_m，最后一列对应 d，$d\neq$null. $\boldsymbol{M}_r=(r_{i,j})$，$j=1,2,\cdots,m$.

$$r_{i,j}=\begin{cases}\text{null} & c_i=\text{null}\\ 0 & c_i(u_j)\neq c_i(u'),\ c_i\neq\text{null}\\ 1 & c_i(u_j)=c_i(u'),\ c_i\neq\text{null}\end{cases}$$

$\boldsymbol{M}_r$ 最后一列 $r_{i,m+1}=d_i$. 定义 $\boldsymbol{M}_r$ 为 $u'(a_j)$的填充矩阵.

通过填充矩阵可以找到样本 u'中缺失值最有可能的值.

定理 1：二进制可辨矩阵的列中 1 的个数表示此列能区分开对象对的个数，此值越大，对应列的重要性越高[5].

定理 2：二进制可辨矩阵中 1 表示对应的属性可以区分开对应的对象对，0 表示对应属性不能区分开对应的对象对，设决策表 T 的二进制可辨矩阵 $\boldsymbol{M}_t$ 中所有属性组成的集合为 R,属性集 $R'\subseteq R$ 为 T 的约简，R'和 R 中属性在 $\boldsymbol{M}_t$ 中属性值 1 对应的对象对完全相同表示 R'没有改变 T 的不可分辨关系[5].

2　基于二进制可辨矩阵的数据补齐算法

对于待填补的属性值，与其有因果关系的属性值缺失过多，待填补的属性值是无法确定的，故对于这种缺失的属性值是没有必要填充的，可以在算法 2 设定一个阈值，根据定义 3 中的缺失比舍弃缺失值过多的样本. 算法 2 首先把待填充样本 u_i 的待填充属性值对应属性 a_j 作为决策属性，利用算法 1 求出约简；再利用约简通过算法 3 建立填充矩阵预测 a_j 的值，算法 2 中求出 2 个约简使得算法可以在数据缺失严重的情况下更好的利用已知属性值来预测未知属性值.

算法 1　求约简的算法

输入：T_1

输出：一个约简 R_1

（1）把 a_j 作为 T_1 的决策属性，对 T_1 构建二进制可辨矩阵 $\boldsymbol{M}_t$；

（2）根据定理 1 找出 $\boldsymbol{M}_t$ 中属性重要性最大的属性 c_i；

（3）在 $R_1\leftarrow\{c_i\}$，$\boldsymbol{M}_t\leftarrow\boldsymbol{M}_t$ 中去掉 c_i 列和 c_i 中 1 对应的行；

（4）if $\boldsymbol{M}_t\neq$null then goto step2. 由定理 2 易得，$\boldsymbol{M}_t$ 为空表示 R_1 没有改变 T_1 的不可分辨关系；

（5）输出 R_1 即为决策表 T_1 的一个约简.

算法 1 求出的 R_1 是 T_1 的近似最大约简,把算法 1 中的第 2 步改为“找出 $\boldsymbol{M}_t$ 中重要性最小的属性 c_i”,则求出 T_1 的近似最小约简 R_2. 由于 R_2 中可能存在冗余属性，故还需要在第 4 步和第 5 步之间利用文献[6]中的去冗余算法去掉 R_2 中可能存在的冗余属性.

算法2 数据补齐算法

输入：T_2

输出：不含缺失数据的样本集

（1）对 T_2 按属性重要性由大到小排序；

（2）把待填充样本 u_i 的缺失属性值对应属性 a_j 作为决策属性，利用算法1求出近似最大约简 R_1 和近似最小约简 R_2；

（3）if u_i-R_2 的缺失比为 0 then {$R \leftarrow R_2$, goto Step7}；

（4）if u_i-R_1 的缺失比为 0 then {$R \leftarrow R_1$, goto Step7}；

（5）if 2个缺失比均大于1 then {删除该样本, goto Step8}；

（6）取小的缺失比对应约简 R_i，设 a_j 中缺失属性值对应属性集合为 A'，$R \leftarrow \{R_i - A'\}$；

（7）调用算法3得到填充值；

（8）$i=i+1$; if $u_i(a_j)$=null then next u_i else goto Step3, until T_2 中样本被遍历一遍；

（9）$j=j+1$; if $u_i(a_j)$=null then next u_i else goto Setp2, until T_2 被填充完毕.

算法3 填充值算法

输入：一个待填充样本，决策表 T

输出：填充完一个待填充值 a_j 的待填充样本

（1）对待填充样本 u' 构建 u' (a_j) 的填充矩阵 $\boldsymbol{M}_r$；

（2）算法2中得到的属性集 $R \cup D$ 与 $\boldsymbol{M}_r$ 中的属性集$\{C,D\}$取交集，形成新的 $\boldsymbol{M}_r$；

（3）计算 $\boldsymbol{M}_r$ 中每行中1的和，并求出其最大值对应的行，取这些行在决策表 T 中样本，统计这些样本中属性 a_j 对应的数值相同，且数量最多的数值作为所求的待填充值.

例2. 设 $T=(U, A)$, $A=\{a_1, a_2, a_3, a_4\}$，待填充样本 u_x=(2, 3, null, 5)；设算法3中步（3）得到最大值对应 m、p、q 行，对应 T 中样本为 u_m、u_p、u_q，并且 $u_m(a_3)=2$、$u_p(a_3)=3$、$u_q(a_3)=2$，则填充值为2，u_x=(2,3,2,5).

$\boldsymbol{M}_r$ 行中1的个数与待填充属性的相似程度成正比. 算法3使用决策表 T 的全部样本中与缺失值有因果关系的值预测缺失值，比仅使用完备样本进行预测的准确率高.

3　试　验

试验选用UCI机器学习数据库中的7个数据集，利用高斯正态分布随机函数，把每个数据集去掉一些数据使含缺失数据样本约占到总样本数的50%. 填补值的正确率=预测值正确的样本数量/含缺失数据样本总数. 比较算法采用文献[7]中的信息增益模型.

表1　填补值正确率

Tab.1　Accuracy of data complement　%

算法	数据集						
	1	2	3	4	5	6	7
本文算法	93	84	68	74	91	76	92
信息增益模型	79	68	51	56	72	57	71

注：表1中各数据集为: 1. Protein localization Sites Database; 2. Glass Identification Database; 3. Heart Disease Database; 4. Australian Credit Approval Database(1); 5. BUPA Liver Disorders Database; 6. Pima Indians Diabetes Database(1); 7. Australian Credit Approval Database(2).

设决策表为 n 行 m 列（若以简化决策表为基础，则设简化后决策表为 n 行 m 列），则求约简的算法的时间复杂度为 O（$mn2$），数据补齐算法的时间复杂度 O（$m3n3$），填充值算法时间复杂度为 O(mn).

4　结　论

本文提出基于Rough集的方法填补缺失数据，定义缺失比把已知值与缺失值的关系量化，设计的填充矩阵极大限度地使用了 T 中可以利用的信息. 通过与同类算法的实验比较，表明本文方法有较好的效果.

参　考　文　献:

[1] Troyanskaya O, Cantor M,Sherlock G, et al. Missing value estimation methods for DNA microarrays[J]. Bioinformatics, 2001, 17（6）: 520—525.

[2] Kantardzic M. Data mining concepts, models, methods and algorithms[M]. Piscataway: Wiley-Interscience, 2003.

[3] Nakata M, Sakai H. Rough sets handling missing values probabilistically interpreted[C]// Slezak D, Wang G, Szczuka M, et al. Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing. Berlin, Germany: Springer-Verlag, 2005: 325—334.

[4] 支天云，苗夺谦. 二进制可辨矩阵的变换及高效属性约简算法的构造[J]. 计算机科学, 2002, 29（2）: 140—142.

[5] 王　磊，王希雷，马永军. 基于二进制可辨矩阵的贪婪算法[J]. 计算机应用研究，2006, 23（增刊）: 92—93.

[6] 王希雷，马永军. 一种基于二进制可辨矩阵的属性约简算法[J]. 天津科技大学学报, 2005, 20（2）: 54—56.

[7] 刘　鹏，雷　蕾，张雪凤. 缺失数据处理方法的比较研究[J]. 计算机科学, 2004, 31（10）: 155—156.