DOI:10.13656/j.cnki.gxkx.2013.04.018

广西科学 Guangxi Sciences 2013,20(4):341～344

缺失数据的学生能力测评方法研究*

Study on the Ability Evaluation Method of Students based on Missing Data

麦宏元[1,2]
MAI Hong-yuan[1,2]

(1.广西大学数学与信息科学学院,广西南宁 530004;2.广西电力职业技术学院,广西南宁 530007)
(1. College of Mathematics and Information Science, Guangxi University, Nanning, Guangxi, 530004, China; 2. Guangxi Electric Power Institute of Vocational Training, Nanning, Guangxi, 530007, China)

摘要:建立数据缺失情况下的学生能力测评系统,并在新的属性集覆盖率和属性重要性等概念的基础上,提出该测评系统的属性约简和规则提取方法,最后用实例验证方法的有效性.

关键词:能力测评系统 粗糙集 属性的重要性 属性约简 规则提取
中图法分类号:O225,TP301 **文献标识码**:A **文章编号**:1005-9164(2013)04-0341-04

Abstract: The evaluation system of student ability is built in missing data, and the new concepts of coverage of the attribute set and the attribute importance are presented. Then the algorithms for attribute reduction and rule extraction are proposed. Finally, the effectiveness of this method is validated by an example.

Key words: capacity evaluation system, rough sets, attribute importance, attribute reduction, rule extraction

随着高等职业教育招生规模的不断扩大,各招生单位都面临一个急需要解决的问题:提高毕业学生的就业率.而就业率的高低在很大程度上由学生的能力来决定.因此,建立一种测评学生能力的模型和方法很有必要.经典粗糙集理论只能处理数据完整的测评系统,其等价关系是对属性集进行分类的基础,但是,在对学生能力进行实际测评时,经常遇到收集数据不完整或数据有缺失的情形,而当属性集中的数据存在缺失时,就不能再利用等价关系对属性集进行分类.已报道的能对数据有缺失的属性集进行划分的方法有:相似关系[1]、非对称相似关系、量化容差关系、限制容差关系、赋值容差关系[2]等.本文在处理数据有缺失的学生能力测评问题时,尝试利用相似关系进行分类和约简.在相似关系下,通过对各属性重要性的比较和度量,对数据有缺失的学生能力测评系统进行属性约简,再对约简所获得的测评系统进行规则的提取,得出评价模式和测评方法.由于对存在缺失值的学生能力测评系统进行属性约简和规则的提取,必须在上近似和下近似下才能进行,因此,还要建立该系统的上、下近似集.

收稿日期:2013-01-07
修回日期:2013-03-15

作者简介:麦宏元(1965-),男,硕士,讲师,主要从事预测与决策、高等数学的教学和研究。

*广西自然科学基金项目(桂科自 0991027)资助。

1 数据有缺失的测评问题模型的建立

1.1 学生能力测评系统

定义 1.1 设 (U,AT,F) 是一个学生能力测评系统,其中 U 为对象集,AT 为属性集,F 为 U 与 AT 之间的关系集,$f(x,a)(x\in U,a\in AT)$ 表示对象 x 在属性 a 下的取值.若存在一个 $x\in U,a\in AT$,$f(x,a)$ 未知(记作 $f(x,a)=*$),则称测评系统 (U,AT,F) 是数据有缺失的学生能力测评系统,即不完备信息系统.

如果 $AT=C\cup D,C\cap D=\varnothing$,则称测评系统 (U,AT,F) 为数据有缺失的学生能力决策测评系统,其中 C 为条件属性集,D 为决策属性集.

1.2 测评系统的相似关系

定义 1.2[3] 设 (U,AT,F) 是数据有缺失的学生

能力测评系统，$\forall A \subseteq AT$，记

$$R_A = \{(y,x) \in U^2 \mid \forall a \in A, f(y,a) = f(x,a) \vee f(y,a) = * \vee f(x,a) = *\},$$

并称 R_A 为数据有缺失的学生能力决策测评系统的相似关系. 记 $S_A(x) = \{y \in U \mid (x,y) \in R_A\}$，$S_A(x)$ 表示 x 的相似类.

1.3 测评系统的粗糙集模型

定义 1.3[3,4] 设 (U,AT,F) 是数据有缺失的学生能力测评系统，对于 $\forall A \subseteq AT, X \subseteq U$，$X$ 的上近似与下近似为

$$\bar{R}_A(X) = \{x \in U \mid S_A(x) \cap X \neq \varnothing\}, \underline{R}_A(X) = \{x \in U \mid S_A(x) \subseteq X\}.$$

其中 $S_A(x)$ 表示 x 的相似类，$S_A(x) = \{y \in U \mid (x,y) \in R_A\}$.

1.4 测评系统属性重要性

定义 1.4 设 (U,AT,F) 是一个数据有缺失的学生能力测评系统，$U/\{d\} = \{D_1, D_2, \cdots, D_m\}$，$A \subseteq C$，记

$$\eta_A = \frac{1}{|\operatorname{card}(U)|^2} \sum_{i=1}^{m} \operatorname{card}(\bar{R}_A(D_i)),$$

且 η_A 称为 A 的覆盖率.

定义 1.5 设 (U,AT,F) 是一个数据有缺失的学生能力测评系统，$U/\{d\} = \{D_1, D_2, \cdots, D_m\}$，$A \subseteq AT$，则对任意的条件属性 $a \in AT \backslash A$，其属性的重要性定义为

$$\rho(a,A) = \eta_A - \eta_{A \cup \{a\}}.$$

定理 1.1 设 (U,AT,F) 是一个数据有缺失的学生能力测评系统，$A_1, A_2 \subseteq AT$，若 $U/R_{A_1}^{\geqslant} \subseteq U/R_{A_2}^{\geqslant}$，则有 $\eta_{A_1} \leqslant \eta_{A_2}$.

证明 因为 $U/R_{A_1}^{\geqslant} \subseteq U/R_{A_2}^{\geqslant}$，即 $\forall x \in U$，有 $S_{A_1}(x) \subseteq S_{A_2}(x)$，又由于 $\bar{R}_A(X) = \{x \in U \mid S_A(x) \cap X \neq \varnothing\}$，因此若假设 $U/\{d\} = \{D_1, D_2, \cdots, D_m\}$，那么对于 $\forall D_i$，$\bar{R}_{A_1}(D_i) \subseteq \bar{R}_{A_2}(D_i)$，即 $\sum_{i=1}^{m} \operatorname{card}(\bar{R}_{A_1}(D_i)) \leqslant \sum_{i=1}^{m} \operatorname{card}(\bar{R}_{A_2}(D_i))$，又由于 $\eta_A = \frac{1}{|\operatorname{card}(U)|^2} \sum_{i=1}^{m} \operatorname{card}(\bar{R}_A(D_i))$，故 $\eta_{A_1} \leqslant \eta_{A_2}$.

定理 1.2 设 (U,AT,F) 是一个数据有缺失的学生能力测评系统，$A_1, A_2 \subseteq AT$，若 $U/R_{A_1}^{\geqslant} = U/R_{A_2}^{\geqslant}$，则有 $\eta_{A_1} = \eta_{A_2}$.

证明 因为 $U/R_{A_1}^{\geqslant} = U/R_{A_2}^{\geqslant}$，即 $\forall x \in U$，有 $S_{A_1}(x) = S_{A_2}(x)$，故 $\forall D_i, \bar{R}_{A_1}(D_i) = \bar{R}_{A_2}(D_i)$，所以 $\eta_{A_1} = \eta_{A_2}$.

定理 1.3 设 (U,AT,F) 是一个数据有缺失的学生能力测评系统，属性集 $A_1, A_2 \subseteq AT$，对于任意 $a \in A_1 \backslash A_2$，令 $U/\{d\} = \{D_1, D_2, \cdots, D_m\}$，若 $U/R_{A_1}^{\geqslant} \subseteq U/R_{A_2}^{\geqslant}$，且 $\rho(a,A_2) = 0$，则有 $\forall D_i, \bar{R}_{A_1}(D_i) = \bar{R}_{A_2}(D_i)$.

证明 因为 $U/R_{A_1}^{\geqslant} \subseteq U/R_{A_2}^{\geqslant}$，即 $\forall x \in U$，有 $S_{A_1}(x) \subseteq S_{A_2}(x)$，又因为 $\rho(a,A_2) = 0$，而 $\rho(a,A) = \eta_A - \eta_{A \cup \{a\}}$，$\eta_A = \frac{1}{|\operatorname{card}(U)|^2} \sum_{i=1}^{m} \operatorname{card}(\bar{R}_A(D_i))$，因此 $\eta_{A_1} = \eta_{A_2}$，所以 $\bar{R}_{A_1}(D_i) = \bar{R}_{A_2}(D_i)$.

定义 1.6 设 $\rho(a,A)$ 是属性 $a \in AT \backslash A$ 的重要性，$\rho(b,A)$ 是属性 $b \in AT \backslash A$ 的重要性.

(1) 若 $\rho(a,A) > \rho(b,A)$，则属性 $a \in AT \backslash A$ 比属性 $b \in AT \backslash A$ 的重要性大；

(2) 若 $\rho(a,A) < \rho(b,A)$，则属性 $a \in AT \backslash A$ 比属性 $b \in AT \backslash A$ 的重要性小；

(3) 若 $\rho(a,A) = \rho(b,A)$，则属性 $a \in AT \backslash A$ 与属性 $b \in AT \backslash A$ 同等重要.

定义 1.7 设 $\alpha (0 < \alpha < 1)$ 为某一给定的阈值，$\rho(a,A)$ 是属性 $a \in AT \backslash A$ 的重要性.

(1) 若 $\rho(a,A) > \alpha$，则属性 $a \in AT \backslash A$ 的重要性大，是重要属性；

(2) 若 $\rho(a,A) \leqslant \alpha$，则属性 $a \in AT \backslash A$ 的重要性小，是冗余属性.

2 数据有缺失的测评问题的属性约简和规则提取方法

2.1 属性约简方法

定义 2.1[5] 设 (U,AT,F) 是一个数据有缺失的学生能力测评系统，R_B 是 U 上的相似关系，$B \subseteq A$，令 $U/\{d\} = \{D_1, D_2, \cdots, D_m\}$，如果满足以下的条件：

(1) $\forall D_i, \bar{R}_B(D_i) = \bar{R}_{AT}(D_i)$，

(2) $\forall a \in A/B, \forall D_i, \bar{R}_{B-\{a\}}(D_i) \neq \bar{R}_{AT}(D_i)$，

则称属性集 B 是 (U,AT,F) 的一个属性约简.

定义 2.2[5] 设 (U,AT,F) 是一个数据有缺失的学生能力测评系统，且属性集 $a \in B \subseteq AT$，令 $U/\{d\} = \{D_1, D_2, \cdots, D_m\}$，对 $\forall D_i$，如果 $\bar{R}_B(D_i) = \bar{R}_{B-\{a\}}(D_i)$，则称 a 是 B 中不必要的属性，否则称 a 是 B 中必要的属性. 如果属性集 B 中的任意属性 a 在 B 中都是必要的，则称属性集 B 是独立的，否则称为相依的. AT 中所有必要属性组成的集合，称为属性集 AT 的核，记为 $\operatorname{Core}(AT)$. 因此，$\operatorname{Core}(AT) = \{a \in$

$AT \mid \rho(AT,a) > 0\}$.

定义 2.3 设(U,AT,F)是一个数据有缺失的学生能力测评系统，$B \subseteq AT$，$\forall a \in AT - B$，若属性集B中的重要性$\rho(a,B) = 0$，则称属性a相对于属性集B为不重要的属性，否则称为重要的属性.

定理 2.1 设(U,AT,F)是一个数据有缺失的学生能力测评系统，且$B \subseteq A \subset AT$，$\forall a \in AT - A$，若$\rho(a,B) = 0$，则$\rho(a,A) = 0$.

证明 因为当$B \subseteq A$时，有$\forall x \in U, S_B(x) \supseteq S_A(x)$，又由于$\forall a \in AT - A, \rho(a,B) = 0$，即$\rho(a,B) = \eta_B - \eta_{B \cup \{a\}}$，则$\eta_B = \eta_{B \cup \{a\}}$，故对$\forall D_i, \bar{R}_B(D_i) = \bar{R}_{B \cup \{a\}}(D_i)$，因此，对$\forall x \in U, D_i, \bar{R}_A(D_i) = \bar{R}_{A \cup \{a\}}(D_i)$，所以$\rho(a,A) = 0$.

定理 2.2 设(U,AT,F)是一个数据有缺失的学生能力测评系统，且$B \subseteq AT$，若$\rho(AT,B) = 0$且$\forall a \in B$，有$\rho(AT,B - \{a\}) > 0$，则属性集B是测评系统的一个约简.

证明 因为$\rho(AT,B) = 0$，故$\eta_B = \eta_{B \cup \{AT\}}$，即对$\forall D_i, \bar{R}_B(D_i) = \bar{R}_{B \cup \{a\}}(D_i)$. 又因为对$\forall a \in B$，有$\rho(AT,B - \{a\}) > 0$，故对$\forall D_i, \bar{R}_{AT}(D_i) \neq \bar{R}_{B-\{a\}}(D_i)$，则属性集$B$是测评系统的一个约简.

根据前面的理论，容易求出缺失测评系统(U,AT,F)属性集AT的核集$\mathrm{Core}(AT)$. 又由于核集是唯一的，而且是任何约简的子集. 因此，选择核集作为启发式算法的起点，逐步删除$AT - \mathrm{Core}(AT)$中相对于核集不重要的属性，然后再逐步添加$A - \mathrm{Core}(A)$相对于核集最重要的属性，直到$\rho(AT,B) = 0$为止，则属性集B是缺失测评系统的一个约简. 算法步骤如下：

输入：缺失测评系统(U,AT,F).

输出：缺失测评系统的一个约简B.

步骤 1 求出核集. 对$\forall a \in AT$，计算$\rho(a,AT)$，若$\rho(a,AT) > 0$，则$\mathrm{Core}(AT) = \{a \in AT \mid \rho(AT,a) > 0\}$，且令$\mathrm{Core}(AT) = B$.

步骤 2 若$\rho(AT,B) = 0$，则转步骤 4，否则转步骤 3.

步骤 3 $\forall a \in AT - B$，计算$\rho(a,B)$. 若$\rho(a,B) = 0$，则删除属性a，否则选择a满足$\max\{\rho(a,B)\}$. 令$B = B \cup \{a\}$，重复步骤 2.

步骤 4 若对$\forall a \in B$，有$\rho(AT,B - \{a\}) > 0$，则转步骤 5.

步骤 5 输出算法约简B.

2.2 规则提取方法

定理 2.3 设$(U,AT \cup d,F)$是一个数据有缺失的学生能力测评系统，如果属性B为约简，则由$f(y,a) = f(x,a), \forall a \in AT \Rightarrow f(y,d) = f(x,d)$可以化简为$f(y,a) = f(x,a), \forall a \in B \Rightarrow f(y,d) = f(x,d)$.

定理 2.4 设$(U,AT \cup d,F)$是一个数据有缺失的学生能力测评系统，如果属性B为约简，若$\exists a \in B, f(y,a) \neq f(x,a)$可推导出$f(y,d) \neq f(x,d)$. 此时的规则为否定的决策规则.

3 实例分析

对如表 1 所示的评价系统进行相关计算.

表 1 有缺失值的学生能力评价系统

Table 1 Evaluation system of student ability with missing data

U	c_1	c_2	c_3	D
1	3	3	2	3
2	1	1	1	1
3	0	*	1	1
4	3	3	2	2
5	1	1	*	1
6	*	3	2	2

注：c_1表示理论课成绩，c_2表示实训课成绩，c_3表示自学能力.

Note: C_1, Results of theoretical lessons; C_2, Results of practice lessons; C_3, Self-learning ability.

(1)计算各个对象的相似类.

$S_{AT}(1) = \{1,4,6\}, S_{AT}(2) = \{2,5\}, S_{AT}(3) = \{3\}, S_{AT}(4) = \{1,4,6\}, S_{AT}(5) = \{2,5\}, S_{AT}(6) = \{1,4,6\}, U/R_D = \{\{1\},\{2,3,5\},\{4,6\}\}$.

(2) 计算各个对象在条件属性下关于决策属性分类的上近似.

令$D_1 = \{1\}$，则$\bar{R}_{AT}(D_1) = \{1,4,6\}$. 令$D_2 = \{2,3,5\}$，则$\bar{R}_{AT}(D_2) = \{2,3,5\}$. 令$D_3 = \{4,6\}$，则$\bar{R}_{AT}(D_3) = \{1,4,6\}$.

(3) 求出缺失测评系统的属性约简.

由于$\rho(AT,a) = \eta_a - \eta_{AT}$，而$\eta_{AT} = \frac{1}{|\mathrm{card}(U)|^2}\sum_{i=1}^{m}\mathrm{card}(\bar{R}_{AT}(D_i))$. 故需要计算各个条件属性下对象的分类：即$S_{c_1}(1) = \{1,4,6\}, S_{c_1}(2) = \{2,5,6\}, S_{c_1}(3) = \{3,6\}, S_{c_1}(4) = \{1,4,6\}, S_{c_1}(5) = \{2,5,6\}, S_{c_1}(6) = \{1,2,3,4,5,6\}$,

$S_{c_2}(1) = \{1,3,4,6\}, S_{c_2}(2) = \{2,3,5\}, S_{c_2}(3) = \{1,2,3,4,5,6\}, S_{c_2}(4) = \{1,3,4,6\}, S_{c_2}(5) = \{2,3,5\}, S_{c_2}(6) = \{1,3,4,6\}$,

$S_{c_3}(1) = \{1,4,5,6\}, S_{c_3}(2) = \{2,3,5\}, S_{c_3}(3) = \{2,3,5\}, S_{c_3}(4) = \{1,4,5,6\}, S_{c_3}(5) = \{1,2,3,$

$4,5,6\}, S_{c_3}(6) = \{1,4,5,6\}$.

令 $D_1 = \{1\}$,**则** $\bar{R}_{c_1}(D_1) = \{1,4,6\}, \bar{R}_{c_2}(D_1) = \{1,3,4,6\}, \bar{R}_{c_3}(D_1) = \{1,4,5,6\}$.

令 $D_2 = \{2,3,5\}$, **则** $\bar{R}_{c_1}(D_2) = \{2,3,5,6\}, \bar{R}_{c_2}(D_2) = \{1,2,3,4,5,6\}, \bar{R}_{c_3}(D_2) = \{1,2,3,4,5,6\}$.

令 $D_3 = \{4,6\}$,**则** $\bar{R}_{c_1}(D_3) = \{1,2,3,4,5,6\}, \bar{R}_{c_2}(D_3) = \{1,3,4,6\}, \bar{R}_{c_3}(D_3) = \{1,4,5,6\}$.

又由于 $\eta_{AT} = \frac{9}{36}, \eta_{C_1} = \frac{13}{36}, \eta_{C_2} = \frac{14}{36}, \eta_{C_3} = \frac{14}{36}$,

故 $\rho(AT,c_1) = \frac{4}{36} > 0, \rho(AT,c_2) = \frac{5}{36} > 0, \rho(AT,c_3) = \frac{5}{36} > 0$. **即** $\mathrm{Core}(AT) = \{c_1,c_2,c_3\} = AT$.

由于 $\rho(AT,AT) = 0$,**故** $AT = \{c_1,c_2,c_3\}$ **是表1的缺失测评系统的约简.**

(4) **缺失测评系统的规则提取.**

$f(c_1) = 3 \wedge f(c_2) = 3 \wedge f(c_3) = 2 \rightarrow f(D) = 3$;

$f(c_1) = 1 \wedge f(c_2) = 1 \wedge f(c_3) = 1 \rightarrow f(D) = 1$;

$f(c_1) = 0 \wedge f(c_2) = * \wedge f(c_3) = 1 \rightarrow f(D) = 1$;

$f(c_1) = 3 \wedge f(c_2) = 3 \wedge f(c_3) = 2 \rightarrow f(D) = 2$;

$f(c_1) = 1 \wedge f(c_2) = 1 \wedge f(c_3) = * \rightarrow f(D) = 1$;

$f(c_1) = * \wedge f(c_2) = 3 \wedge f(c_3) = 2 \rightarrow f(D) = 2$.

简化得

$f(c_1) = 1 \wedge f(c_2) = 1 \rightarrow f(D) = 1; f(c_1) = 0 \wedge f(c_3) = 1 \rightarrow f(D) = 1$;

$f(c_2) = 3 \wedge f(c_3) = 2 \rightarrow f(D) = 2; f(c_1) = 3 \wedge f(c_2) = 3 \wedge f(c_3) = 2 \rightarrow f(D) = 3$.

综上所述,本文提出的方法是有效的.

参考文献:

[1] **章正辉,戴小鹏,熊大红,等. 不完备决策表的启发式知识约简算法研究**[J]. **计算机与现代化**,2010(03):170-172.

[2] **梅良才. 不完备信息系统中基于粗糙集的多属性决策问题研究**[D]. **南宁:广西大学**,2010.

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

[4] **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.

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

(责任编辑:尹　闯)

广西区科协第七次代表大会召开

新闻时间 2013-10-28

10月25日～26日,广西壮族自治区科学技术协会第七次代表大会在南宁举行。自治区党委书记彭清华、中国科协副主席程东红出席会议并讲话。大会选举中国工程院院士郑皆连任广西壮族自治区科协新一届主席。

彭清华指出,希望全区科技工作者立足基本区情,努力将个人追求与"中国梦"及广西发展战略深度融合;树立创新趋动发展理念,善走产学研相结合之路,加快科技成果向现实生产力转化;坚持兼容并举,积极发掘和培养科技新生力量。

程东红代表中国科协对大会的召开表示热烈祝贺。她希望广西壮族自治区科协以伟大的中国梦凝聚智慧力量,充分调动科技工作者的积极性、创造性;发挥学会在创新体系中的重要作用,在实施创新驱动发展战略中作出新贡献;肩负提高全民科学素质的崇高使命,为建设创新型国家筑牢基础;着力促进人才成长,形成各类创新人才竞相涌现的局面。

大会听取并审议通过了广西壮族自治区科协第六次委员会工作报告,选举产生了新一届委员会主席、副主席、常委。大会还对"第十二届广西青年科技奖"、"2008～2013年广西科协系统先进集体和先进工作者"进行了表彰。

摘自《中国科学报》