Volume 3, Issue 9, September 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Improvement of Method Handling Missing Values in Incomplete Information System

Hung Quoc Nguyen*
Department of Information and International Cooperation in Research Institute for Aquaculture No 3, Vietnam

Duc Thuan Nguyen
Department of Information Systems, Nha Trang University, Vietnam

Abstract— Many methods have been proposed to process missing data for information system. In the paper, we modified an algorithm to handle missing value based on covering rough sets model previously reported by Dai Dai and Jianpeng Wang proposed to transform an incomplete information system into a complete information system. The experimental results show that the new version of algorithm is efficient.

Keywords— Rough sets, Covering rough sets, Incomplete Information system

I. INTRODUCTION

Rough set theory was first established by Polish mathematician Z.Pawlak as a formal tool for modelling and processing the incomplete information in information system [9]. Unfortunately, incomplete information systems (IIS) can be seen everywhere in actual world [4], [5], [6], [7], [8]. There are usually two methods in rough set theory (RST) to deal with an incomplete information system [4]. The first is an indirect method that transforms an incomplete information system to a complete one in some ways (e.g probability statistical methods usually). It is called data reparation also [6]. The second is a direct method that extends the concepts of the classical RST to process incomplete information [2], [4], [7]. In this paper, we modified an algorithm created by Dai Dai and Jianpeng Wang [2] that was proposed to handle missing values for incomplete information system. It inherits the merit of the other extensions of the classical rough set theory and avoid their limitation. The paper is organized as follows. Some preliminary concepts are briefly recalled in Section II. In Section III, we present Dai Dai & Jianpeng Wang's estimating unknown values method. Then, shortcomings of Dai Dai & Jianpeng Wang's method and the solutions to the shortcomings in Section IV. In Section V, we present two illustrative examples from UCI databases. Section VI concludes this paper.

II. PRELIMINARIES

A. Covering Rough Sets Model

Definition 1 Let A be a set, the set family $\{A_i \subseteq A \mid i = 1, n\}$ is a partition of A, where

- $A_i \neq \emptyset$, $\forall i = 1..n$
- $A_i \cap A_j = \emptyset$ $\forall i, j=1..n, i \neq j$
- $\bigcup_{i=1}^{n} A_i = A$

Definition 2 Let *U* be a domain, *C* is a family of none empty subsets of *U*. If $\cup C = U$, then C is called a covering of *U*.
It can be seen that a partition of *U* is a covering of U. Consequently, the covering is an extra notion of the partition.

Definition 3 Let *U* be a domain, *C* is a covering of *U*. The ordered pair <*U*, *C*> is called a covering approximation space.

Definition 4 Let <U, *C*> be a covering approximation space, x∈ U, then the set family

$$\{K \in C \mid x \in K \wedge (\forall S \in C \wedge x \in S \wedge S \subseteq K \Rightarrow K = S)\}$$

is called the minimal description of x and denoted by md(x).

Definition 5 Let <U, *C*> be a covering approximation space, for any X ⊆U, the covering upper and lower approximation set families of X are respectively defined as

$$\overline{C}(X) = \{x \in U \mid \bigcap md(x) \cap X \neq \emptyset\}, \quad \underline{C}(X) = \{x \in U \mid \bigcap md(x) \subseteq X\}.$$

B. Covering Reducts

Definition 6 Let <*U*, *C*> be a covering approximation space, K ∈ *C*, if K is the union of sets in *C* –{K}, then K is called a reducible element of *C*, otherwise K is called an irreducible element of *C*.

Definition 7 Let <*U, C*> be a covering approximation space, if all element in *C* are irreducible elements, then *C* is called an irreducible covering, otherwise *C* is called a reducible covering.

Definition 8 Let <*U, C*> be a covering approximation space, the irreducible covering after removing all the reducible element of *C* is called the reduct of *C*, and denoted by reduct(*C*).

C. Incomplete Information System (IIS)

Information system S is a tuple S = <*U, A, V, f*>, where

- U is a nonempty finite set of objects called the universe of discourse;
- A is a nonempty finite set of attributes;
- $V = \bigcup_{a \in A} V_a$ where V_a = Dom (a) ≠ Ø, $|V_a| < \infty$;
- f : U x A $\rightarrow$ V
  (x, a) $|\rightarrow$ v ∈ V_a;
- if ∃c∈A, x∈U, f(x,c) = "*" (f(x,c) is not determined), then S is called incomplete information system (IIS), otherwise it is complete.
- S is called decision system, if A = C∪D, C∩D = Ø, where C is set of condition attributes and D is set of decision attributes. The decision table corresponding with S is denoted by T = <*U, C∪D*>.

Incomplete equivalence classes

Let S = <U, A, V, f> be an information system, for any attribute $a \in A$, each object x ∈U is presented by a corresponding tuple (obj, symbol) as

$$(\text{obj, symbol}) = \begin{cases} (obj, u), if\ f(a) = '*' \\ (obj, u), if\ f(a) \neq '*' \end{cases}$$

For attribute a ∈A, equivalence relation χ_a on U is expressed as:

∀x, y ∈ U : x χ_a y ↔ (f (x, a) = f (y, a)) (f (x, a) = '*') (f (y, a) = '*').

The incomplete equivalence class containing an element x is denoted by

[x] χ_a ={y ∈ U | x χ_a y}.

III. DAI DAI & JIANPENG WANG'S ESTIMATING UNKNOWN VALUES METHOD

Let S = <*U, C∪D, V, f*> be a decision system. It can be seen that

C = {$[x]_{\chi a}$ | a ∈ C, ∀x ∈ U} is a covering of U.

The object unknown value can be estimated in the following two rules [2]:

R1: If x, y ∈$\bigcap md(z)$ and f (x,d) = f (y,d), ∀d ∈D, then all the condition attribute values of x and y can be transformed to the known value. (unknown attribute values of x (y) are replaced by corresponding known attribute values of y (x)).

R2: If x, y ∈$\bigcap md(z)$ and ∃d ∈D: f (x,d) ≠ f (y,d), then the missing values of objects are processed in the following two ways:

(Assuming that x is the missing value object)

W1: Finding z' where x, v ∈ $\bigcap md(z)$ and x, v satisfying R1, then x is evaluated.

W2: Believing that x, y are distinguished, then the unknown values of x are evaluated by values so that x ≠ y.

IV. SHORTCOMINGS OF DAI DAI & JIANPENG WANG'S METHOD AND THE SOLUTIONS TO THE SHORTCOMINGS

To estimate missing values of a object x, Dai Dai & Jianpeng Wang's method is performed in two steps that are expressed respectively as follows:

Step 1 Finding any object y having no missing value in any ∩md(z) which contains object x where x, y ∈ ∩md(z) and f(x,d) = f(y,d), ∀d ∈ D, then the missing control attribute values of x can be transformed to the known values of corresponding attributes of y.The shortcoming is that it can find out more than one object y. To deal with the problem, we can estimate the unknown values of x according to the objects of y which havethe largest occurence frequency.

Step 2 If not finding out any object y according to Step 1, then finding all objects y having no missing value where x, y ∈ ∩md(z), and ∃d ∈ D where f(x,d) ≠ f(y,d). And then the missing values of x are evaluated so that the condition attribute values are not identical to those of y.

The shortcoming is that it? Can does not exist value to estimate the missing value, so that the condition attribute value is not identical. In this case, object x is believed to be in conflict with other objects in the input data set U, so x can be eliminated from U.

Another shortcoming of this method is that if there are exists ∩md(z) which only contain objects missing values (without any complete object), so there can be exist object x ∈∩md(z) where x has no connection with other objects to be able to estimate the missing values of x. To surmount this problem we can remove the object x from U.

In the rest of this paper, we will illustrate the algorithm with a concrete example.

Example 1 Let an incomplete information system as shown in **Table I** (Location, Basement, Fireplace: condition attribute; Value: decision attribute)

TABLE I
AN INCOMPLETE INFORMATION SYSTEM

Objects	Location	Basement	Fireplace	Value
1	good	yes	yes	high
2	bad	*	no	small
3	good	no	*	medium
4	bad	yes	no	medium
5	*	yes	no	medium
6	good	yes	*	small
7	good	yes	no	medium

The algorithm consists of the steps that are expressed respectively as follows:

Step 1: Represent incomplete equivalent classes of all the condition attribute values as follows:

U/{Location}={{(1,c),(3,c),(5,u),(6,c),(7,c)},{(2,c),(4,c),(5,u)}};

U/{Basement}={{(1,c),(2,u),(4,c),(5,c),(6,c),(7,c)}, {(2,u),(3,c)}};

U/{Fireplace}={{(1,c),(3,u),(6,u)},{(2,c),(3,u),(4,c),(5,c),(6,u),(7,c)}};

The all incomplete equivalence classes' union of the condition attributes: {{(1,c),(3,c),(5,u),(6,c),(7,c)},{(2,c),(4,c),(5,u)},{(1,c),(2,u),(4,c),(5,c),(6,c),(7,c),{(2,u),(3,c),{(1,c),(3,u),(6,u)},{(2,c),(3, u),(4,c),(5,c),(6,u),(7,c)}};

Step 2: Computing $\cap md(obj^i)$, i = 1, 2, …, n

$\cap md(1) = \{1, (6,u)\}$

$\cap md(2) = \{(2,u)\}$

$\cap md(3) = \{(3,u)\}$

$\cap md(4) = \{(2,u), 4, (5,u)\}$

$\cap md(5) = \{(5,u)\}$

$\cap md(6) = \{(6,u)\}$

$\cap md(7) = \{(5,u), (6,u), 7\}$

Step 3: Simplifying sets $\cap md(obj^i)$.

If $\cap md(j) \subseteq \cap md(i)$, then remove $\cap md(j)$. In this way, for the results in Step 2, $\cap md(obj^i)$ after the simplifying is obtained as

$\cap md(1) = \{1,(6,u)\}$

$\cap md(3) = \{(3,u)\}$

$\cap md(4) = \{(2,u),4,(5,u)\}$

$\cap md(7) = \{(5,u), (6,u), 7\}$

Step 4: Eveluating the value of the missing attributes.

In $\cap md(1) = \{1,(6,u)\}$, f(1,Value) = high and f(6,Value) = small, so the estimated value of f(6,Fireplace) ≠ f(1,Fireplace) = **yes**; in $\cap md(7) = \{(5,u), (6,u), 7\}$, f(7,Value) = medium and f(6,Value) = small, so the estimated value of f(6,Fireplace) ≠ f(7,Fireplace) = **no**; thus not there not exist value to estimate value for 6th object, then 6th object is deleted from U.

In $\cap md(3) = \{(3,u)\}$, 3rd object contains unknown value but there is not any connection with it, so it can be filled with value. Then 3rd object is removed from U.

In $\cap md(4) = \{(2,u), 4, (5,u)\}$, f(2,Value) = small and f(4,Value) = medium, so the estimated value of f(2,Basement) ≠ f(4,Basement) = **yes**, then f(2, Basement) = **no**.

In $\cap md(4) = \{(2,u), 4, (5,u)\}$, f(4,Value) = f(5,Value) = medium, so the estimated value of f(5,Location) = f(4,Location) = **bad**; in $\cap md(7) = \{(5,u), (6,u),7\}$, f(5,Value) = f(7,Value) = medium, so the estimated value of f(5, Location) = f(7,Location) = **good**; let P(4), P(7) are respectively frequency of 4th and 7th objects in U, (i) if P(4) > P(7) then f(5,Location) = **bad**, (ii)

if P(4) < P(7) then f(5,Location) = **good**, (iii) if P(4) = P(7) then f(5,Location) = **bad** or f(5, Location) = **good**. With the information system as Table 1, the frequencies of 4th and 7th objects are equal at 1, so f(5,Location) = **bad** or f(5,Location) = **good** (corresponding to case (iii)).

TABLE II
THE INFORMATION SYSTEM AFTER ESTIMATING THE MISSING VALUES

Objects	Location	Basement	Fireplace	Value
1	good	yes	yes	high
2	bad	**no**	no	small

4	bad	yes	no	medium
5	**good** or **bad**	yes	no	medium
7	good	yes	no	Medium

V. EXPERIMENTS

The following experiment used Mushroom and Mammographic Mass datasets obtained from UCI machine learning repository. The experiment was executed on Intel(R) Core(TM) i3 CPU 2.27 GHz machine and the software was VC#. The experiment results are listed in **Table III**.

TABLE III
RESULTS OF THE EXPERIMENT USING MUSHROOM AND MAMMOGRAPHIC MASS DATASETS FROM UCI

Dataset Name	**Number of Instances**	**Number of Attributes**	**Number of Missing Values**	**Number of Instances after Experiment**	**Exec Time. (*milliseconds*)**
Mushroom	8124	23	2480	10580	725195
Mammographic Mass	961	6	162	1074	734

VI. CONCLUSIONS

This paper focused on studying the algorithm of Dai Dai & Jianpeng Wang [2] to estimate the missing values in the incomplete information system. In the studying process of this algorithm, we found a few shortcomings and the solution for the problem was also proposed. We will in the future work on algorithms and applications of covering rough sets in data mining and reduction.

REFERENCES

[1] Chen Wu, Enbin Wang, Xibei Yang, *Knowledge Dependency in Expanded Incomplete Information Systems*, IJCSNS International Journal of Computer 292 Science and Network Security, VOL.8 No.5, May (2008).
[2] Dai Dai, Jianpeng Wang, *A New Extracting Rule Algorithm from Incomplete Information System*, International Conference on Intelligent Systems and Knowledge Engineering Proceedings (ISKE-2007), (2007).
[3] Guoyin Wang, *Extension of Rough Set under Incomplete Information Systems*, Journal of Computer Research and Development (in Chinese), 39(10): 1238~1243 (2002).
[4] Kryszkiewicz,M.: *Rough set approach to incomplete information systems Information Sciences*, 39-49, 112 (1998).
[5] Jerzy W. Grzymala-Busse, *Three Approaches to Missing Attribute Values—A Rough Set Perspective*, Accepted for the Workshop on Foundations of Data Mining, associated with the fourth IEEE International Conference on Data Mining, Brighton, UK, November 1–4, (2004).
[6] Jerzy W. Grzymala-Busse, Sachin Siddhaye, *Rough Set Approaches to Rule Induction from Incomplete Data*, Proceedings of the IPMU'2004, the 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Perugia, Italy, vol. 2, 923–930, July 4–9, (2004).
[7] Xuri Yin, Xiuyi Jia, Lin Shang, *A New Extension Model of Rough Sets Under Incomplete Information*, Rough Sets and Knowledge Technology Lecture Notes in Computer Science Volume 4062, pp 141-146, (2006).
[8] Yao, H., Hamilton, H.J., and Butz, C.J., *A Foundational Approach for Mining Itemset Utilities from Databases*, In Proceedings 2004 SIAM International Conference on Data Mining (SIAMDM04), Orlando, FL, April, (2004).
[9] ZdzislawPawlak. *Rough Sets: Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, Norwell, MA, USA, (1992).

