International Journal of Database Theory and Application
Vol.8, No.6 (2015), pp.149-164
http://dx.doi.org/10.14257/ijdta.2015.8.6.14

A Novel Data Filling Algorithm for Incomplete Information System Based on Valued Limited Tolerance Relation

Xiuling Bai, Mingchuan Zhang, Qingtao Wu, Ruijuan Zheng,
Haixia Zhao and Wangyang Wei

College of Information Engineering, Henan University of Science and Technology,
Luoyang, Henan Province, 471023, China
Email: baixiuling@163.com, zhlzmc@163.com, wqt8921@haust.edu.cn, rjwo@163.com, lizhao@163.com, weiwangyang@163.com

Abstract

Due to various reasons, there are generally missing data in datasets. Usually the missing data in these incomplete datasets need to be filled. In this paper, the drawbacks of some existing data filling approaches for incomplete information systems are analyzed based on Rough Set theory. Several similarity relation models are discussed and the Valued Limited Tolerance Relation model is proposed. A data filling algorithm based on the Valued Limited Tolerance Relation model is put forward. This approach makes full use of the similarity of objects and selects the object which is the most similar to the incomplete object. More missing data can be filled scientifically. The experimental results show that this approach is effective.

Keywords: *Rough Set, Information System, Incomplete Data, Data Filling*

1. Introduction

Real-life data are frequently imperfect, erroneous, incomplete, uncertain and vague. Due to the errors in data measuring, data understanding, data registration and *etc.*, it is common that missing data exist in datasets. Usually the missing data in these incomplete datasets need to be filled.

At present, the missing data are tackled mainly through the following approaches [1]. 1) Method of Deleting Objects with Unknown Attribute Values. The method is the simplest, *i.e.*, just delete the objects which have missing attribute values. 2) Method of Treating Missing Attribute Values as Special Values. The method treats "unknown" itself as a new value and treat it in the same way as other values. 3) Method of Filling in the Missing Data according to the Distribution of the Remaining Objects' Attributes. There are several algorithms. In *Mean Completer algorithm*, it substitutes missing values for numerical attributes with the mean value of all values for that attribute. For non-numerical attributes, missing values are substituted by the "mode" value, *i.e.*, the most frequently occurring value among the values for that attribute. *Conditioned Mean Completer algorithm* comes from the *Mean Completer algorithm*, but the mean and mode values are obtained from the objects with the same decision attributes. In *Combinatorial Completer algorithm*, an object is expanded into several objects covering all possible combinations of the object's missing values. *Conditioned Combinatorial Completer algorithm* comes from the above one but the sets of values are conditioned to decision class. Most of these approaches mentioned above are based on the probability statistics. Since the state space of dataset is usually very large, the distribution is difficult to be found. Therefore, traditional statistical techniques are not optimal approaches.

ISSN: 2005-4270 IJDTA

Rough Set theory was put forward by Pawlak in 1982 [2]. It has been achieved a great success in data mining, data analyzing, malfunction analyzing, knowledge acquiring [3], and others [4]. It is a useful mathematical tool for depicting incompleteness and indetermination. It can effectively analyze and deal with inexact, uncertain or vague knowledge.

Many approaches have been studied to fill in the missing data according to the indiscernibility relation in Rough Set theory. ROUSTIDA algorithm is one of them [5]. The basic idea is that the classification rules generated after filling in missing data should have the support as high as possible. This algorithm is effective, but it doesn't discriminate condition attributes and decision attributes. Therefore, it isn't suitable for decision systems. Some improved algorithms were put forward based on ROUSTIDA algorithm [6-8]. Decision attributes will be priori filled and the potential conflict of decision rules will be eliminated in these algorithms. These algorithms can be used in decision systems. ROUSTIDA algorithm and the improved algorithms are all based on Tolerance Relation model [9-10]. The algorithms are rather simple. However, if there are conflicts of the attribute values in similar objects, the data cannot be filled and other approaches are needed. Some algorithms [11-13] are put forward based on Valued Tolerance Relation model [14]. They can effectively solve the problem mentioned above. However, two objects which have no clearly same attribute values may also wrongly belong to the same tolerance class.

To solve this problem, characteristics of Limited Tolerance Relation and Valued Tolerance Relation are combined in this paper. The Valued Limited Tolerance Relation model is proposed. A data filling algorithm based on the Valued Limited Tolerance Relation model is given.

2. Basic Rough Set Notions

Rough Set theory is an extension of set theory, in which a subject of universe is described by a pair of ordinary sets called low and upper approximations. A key notion in Rough Set theory is the equivalence relation [2]. The equivalence classes are the building blocks for the construction of the low and upper approximations.

Definition 1. Information system is a pair $I =< U, AT >$, where U is a non-empty finite set of objects and called universe. If there are n objects, U can be expressed as $U = \{u_1, u_2, ..., u_n\}$. AT is a non-empty finite set of attributes. If there are m attributes, AT can be expressed as $AT = \{a_1, a_2, ..., a_m\}$. V_a is the domain of the attribute a. If there is any $x \in U$, $a \in AT$, and missing values contained in V_a (* denotes missing value), then I is called an incomplete information system, otherwise it is complete. AT is divided into two non-intersect sets: C is called condition attributes set and D is called decision attributes set. C and D satisfy $AT = C \bigcup D$ and $C \bigcap D = \varnothing$, then the information system is called decision system. Decision system is always appeared as a planar table, so it is called decision table.

Definition 2. Each subset of attributes $A \subseteq AT$ determines a binary Indiscernibility Relation $IND(A)$ as follows: $IND(A) = \{(x, y) \in U \times U \mid \forall a \in A, a(x) = a(y)\}$. $I_A(x)$ will denote the set of objects indiscernible with x by A ($I_A(x) = \{y \in U \mid (x, y) \in IND(A)\}$).

Definition 3. Let $X \subseteq U$, R is an equivalence relation. When X is an union of R basic categories, X is called R definable, otherwise X is called R non-definable. R definable set is also called R exact set, and R non-definable set is also called R rough set.

Definition 4. Let $X \subseteq U$ and $A \subseteq AT$. $\underline{A}X$ is called lower approximation of X iff $\underline{A}X = \{x \in U \mid I_A(x) \subseteq X\}$. $\underline{A}X$ is the set of objects that certainly belong to X. $\overline{A}X$ is

called upper approximation of X iff $\overline{A}X = \{x \in U \mid I_A(x) \bigcap X \neq \varnothing\}$. $\overline{A}X$ is the set of objects that possibly belong to X. $\underline{A}X \subseteq X \subseteq \overline{A}X$.

3. Data Filling Algorithm for Incomplete Information System

3.1. Tolerance Relation and ROUSTIDA Algorithm

ROUSTIDA algorithm was proposed in paper based on Indiscernibility Relation in Rough Set theory [5]. The basic idea is that the classification rules generated after filling in missing data should have the support as high as possible and be centralized as far as possible. If the support of rules is small and the rules are widely distributed, the information may contain noisy rules. In other words, the aim of this approach is to retain the consistency between the objects with missing values and other similar objects in information systems. The difference between attributes should be as small as possible. The condition attributes and the decision attributes are not distinguished in this algorithm, so it is mainly applied to information systems without decision.

It is necessary for incomplete information systems to define the similarity of object. Tolerance Relation [9] is the basic relation in ROUSTIDA algorithm.

Definition 5. Given an incomplete information system $S = (U, A, V, f)$, where $a = U \bigcup \{d\}$, C is the set of condition attributes and d is decision attribute. For any subset $B \subseteq A$ which has missing attribute values, the Tolerance Relation T can be defined as in (1):

$$\forall_{x,y \in U}(T_B(x,y) \Leftrightarrow \forall_{c_j \in B}(c_j(x) = c_j(y) \vee c_j(x) = * \vee c_j(y) = *)) \quad (1)$$

Discernibility Matrix proposed by Skowron [15-16] in 1992 is one of the basic concepts in Rough Set theory. It concentrates all distinct information of attributes in one matrix. It has always been playing an important role in various analyses and studies based on Rough Set theory. Discernibility Matrix can be defined as in (2).

Definition 6. Given an information system $I = < U, AT >$, where $B \subseteq AT$, Discernibility Matrix is a $n \times n$ square matrix $M(B) = \{M(i,j) \mid_{n*n}, 1 \le i \le n = \mid U \mid$. The unit of the matrix can be defined as in (2) :

$$M(i,j) = \{a \in B : a(u_i) \neq a(u_j), u_i, u_j \in U, i, j = 1, 2, ..., n\} \quad (2)$$

Discernibility Matrix is a symmetric square matrix. The unit of the matrix is an attributes set which indicates the differences of attribute values between two objects. Discernibility Matrix reflects the differences of objects, so it is the basis of ROUSTIDA algorithm.

The definition of Discernibility Matrix mentioned above is only suitable for complete information system. It is modified in paper [5] in order to suitable for incomplete information system.

Definition 7. Given an information system $I = < U, AT >$, Generalized Discernibility Matrix is a $n \times n$ square matrix $M_E(AT) = \{M(i,j) \mid_{n*n}, 1 \le i \le n = \mid U \mid$. The unit of the matrix can be defined as in (3):

$$M_E(i,j) = \{k : (a_k(u_i) \neq a_k(u_j)) \bigcap (a_k(u_i) \neq *) \bigcap a_k(u_j) \neq *), k = 1, 2, ..., m; i, j = 1, 2, ..., n\} \quad (3)$$

The following concepts are defined in ROUSTIDA algorithm.

Definition 8. Given an information system $I = < U, AT >$, the missing attributes set MAS_i of object u_i, indistinguishableness objects set NS_i of object u_i and missing objects set MOS of I can be defined respectively as:

$MAS_i = \{k : a_k(u_i) = *, k = 1, 2, ..., m\}$;

$NS_i = \{j : M_E(i,j) = \varnothing, i \neq j, j = 1, 2, ..., n\}$;

$MOS = \{i : MAS_i \neq \varnothing, i = 1, 2, ..., n\}$.

Supposing an original information system I^0, objects set $\{u_i^0\}$, corresponding Generalized Discernibility Matrix M_E^0, the element of Line *i* Row *j* in matrix is $M_E^0(i, j)$. The missing attributes set and indistinguishableness objects set of object u_i^0 are MAS_i^0 and NS_i^0, respectively. The missing objects set of the information system I^0 is MOS^0. The information system after *r*-th completing analysis is I^r, objects set is $\{u_i^r\}$, corresponding Generalized Discernibility Matrix is M_E^r, the missing attributes set and indistinguishableness objects set of object u_i^r are MAS_i^r and NS_i^r, respectively. The missing objects set of the information system I^r is MOS^r. A data filling algorithm ROUSTIDA is put forward based on the above Generalized Discernibility Matrix in paper [5].

Input: incomplete information system $I^0 = \langle U^0, AT \rangle$

Output: complete information system $I^r = \langle U^r, AT \rangle$

Step 1: Computing original Generalized Discernibility Matrix M_E^0, MAS_i^0 ($i = 1, 2, ..., n$) and MOS^0, let $r = 0$;

Step 2:

2.1 for all $i \in MOS^r$, compute NS_i^r;

2.2 Generating I^{r+1}:

2.2.1 for all $i \notin MOS^r$, let $a_k(u_i^{r+1}) = a_k(u_i^r)$, $k = 1, 2, ..., m$;

2.2.2. for all $i \in MOS^r$, while $k \in MAS_i^r$ do:

2.2.2.1 if $|NS_i^r| = 1$, supposing $j \in NS_i^r$;

if $a_k(u_j^r) = *$, then let $a_k(u_i^{r+1}) = *$; else let $a_k(u_i^{r+1}) = a_k(u_j^r)$;

2.2.2.2 else

(i)if $\exists j_0, j_1 \in NS_i^r$, which satisfy $(a_k(u_{j_0}^r) \neq *) \bigcup (a_k(u_{j_1}^r) \neq *) \bigcup (a_k(u_{j_0}^r) \neq a_k(u_{j_1}^r))$

then let $a_k(u_i^{r+1}) = *$;

(ii) if $\exists j_0 \in NS_i^r$, which satisfies $a_k(u_{j_0}^r) \neq *$, then let $a_k(u_i^{r+1}) = a_k(u_{j_0}^r)$;

(iii) else let $a_k(u_i^{r+1}) = *$;

2.3 if $I^{r+1} = I^r$, then end while and goto Step 3;

else compute M_E^{r+1}, MAS_i^{r+1} and MOS^{r+1}; let $r = r + 1$; goto Step 2;

Step 3: If missing data still exist in the information system, they can be filled by other algorithms, such as Mean Completer and Combinatorial Completer.

Step 4: End.

3.2. Valued Tolerance Relation and Data Filling Algorithm

The ROUSTIDA is rather simple. However, if there are conflicts of the attribute values in similar objects, the data cannot be filled and other approaches are needed.

Instance 1. Given an incomplete information system as in Table 1, where a_1, a_2, a_3 are objects, c_1, c_2, c_3, c_4 are four attributes which values (discrete) range from 0 to 3, "*" denotes the missing attribute value.

Table 1. An Information System

A	c_1	c_2	c_3	c_4
a_1	3	2	1	0
a_2	3	*	1	0
a_3	*	3	*	0

The attribute c_2 of object a_2 needs to be filled. According to Tolerance Relation model, a_1 and a_3 are both similar to a_2, but the attribute values in c_2 of two objects are different. Therefore, the attribute c_2 of object a_2 can't be filled in ROUSTIDA algorithm. Intuitively, a_1 is more similar to a_2 than a_3. Tolerance Relation model cannot measure the degree of two objects' similarity, so the Valued Tolerance Relation model is proposed in paper [14].

Definition 9. Given an incomplete information system $S=(U,A,V,f)$, supposing that the possible values of each attribute are uniformly distributed. For $\forall a_k \in A$, the range is $E_k=\{e_k^1,e_k^2,...,e_k^t\}$. For any object $x_i \in U$, $a_k(x_i)=e_k^i$ probability is $1/|E_k|$. Given two objects $x_i \in U$, $x_j \in U$, their similarity in attribute a_k can be defined as in (4):

$$P_k(i,j)=\begin{cases} 1 & (a_k(x_i)=a_k(x_j))\wedge(a_k(x_i)\neq *)\wedge(a_k(x_j)\neq *) \\ 0 & (a_k(x_i)\neq a_k(x_j))\wedge(a_k(x_i)\neq *)\wedge(a_k(x_j)\neq *) \\ \dfrac{1}{|E_k|} & ((a_k(x_i)=*)\wedge(a_k(x_j)\neq *))\vee((a_k(x_j)=*)\wedge(a_k(x_i)\neq *)) \\ \dfrac{1}{|E_k|^2} & (a_k(x_i)=*)\wedge(a_k(x_j)=*) \end{cases} \tag{4}$$

The Valued Tolerance Relation Matrix is defined as in (5):

$$M_V(i,j)=\begin{cases} 1 & i=j \\ \prod_{a_k\in A} P_k(i,j) & i\neq j \end{cases} \tag{5}$$

A data filling algorithm VTRIDA based on Valued Tolerance Relation model is proposed in paper [11].

If the original Valued Tolerance Relation Matrix is M_V^0, the Valued Tolerance Relation Matrix after *r*-th completing analysis is M_V^r.

Input: incomplete information system $I^0=\langle U^0,AT\rangle$

Output: complete information system $I^r=\langle U^r,AT\rangle$

Step 1: Computing original Valued Tolerance Relation Matrix M_V^0, MAS_i^0 ($i=1,2,...,n$) and MOS^0, let $r=0$;

Step 2:

2.1 Generating I^{r+1}:

2.1.1 for all $i\notin MOS^r$, let $a_k(u_i^{r+1})=a_k(u_i^r)$, $k=1,2,...,m$;

2.1.2 while $i\in MOS^r$ and $T(i,j)>0$, do:

seek j', which satisfy $T(i,j')=\max(T(i,j))$.

firstly, if $\exists j'$, then: $a_k^{r+1}(u_i)=\begin{cases} a_k^r(u_{j'}), & if\ a_k^r\in MAS_i^r \\ a_k^r(u_i), & if\ a_k^r\notin MAS_i^r \end{cases}$ $k=1,2,...,m$

secondly, if j' doesn't exist, then let $a_k(u_i^{r+1}) = a_k(u_i^r)$, $k = 1, 2, ..., m$;

2.2 if $I^{r+1} = I^r$, then endwhile and goto Step 3;

else compute M_V^{r+1}, MAS_i^{r+1} and MOS^{r+1}; let $r = r + 1$; goto Step 2;

Step 3: If missing data still exist in the information system, they can be filled by other algorithms such as Mean Completer and Combinatorial Completer.

Step 4: End.

3.3. Limited Tolerance Relation

Both in the Tolerance Relation model and the Valued Tolerance Relation model, the unknown values "*" are looked upon as equaling to any known attribute values. On this condition, even if two objects have no (or very few) clearly same attribute values, they may also wrongly belong to the same tolerance class.

Instance 2. Given an incomplete information system as in Table 2, where a_1, a_2, a_3 are objects, c_1, c_2, c_3, c_4 are four attributes which values (discrete) range from 0 to 3, "*" denotes the missing attribute value.

Table 2. An Information System

A	c_1	c_2	c_3	c_4
a_1	0	*	*	1
a_2	*	2	3	3
a_3	0	*	*	*

According to Tolerance Relation model and Valued Tolerance Relation model, objects a_2 and a_3 are considered as indiscernible and similar. However, these two objects haven't any equal attribute values and there is just a little probability for their attribute values to be equal to each other.

The attribute c_4 of object a_3 needs to be filled. According to Tolerance Relation model, a_1 and a_2 are both similar to a_3, but the attribute values in c_4 of two objects are different. Therefore, the attribute c_4 of object a_3 can't be filled in ROUSTIDA algorithm.

According to the Valued Tolerance Relation model, Valued Tolerance Relation Matrix is obtained as in Table 3.

1/256>1/1024, the attribute value 3 of object a_2 should be chosen to fill in attribute c_4 of object a_3. But intuitively, a_2 and a_3 have no common attributes, a_1 is more similar to a_3 than a_2.

Table 3. Valued Tolerance Relation Matrix

	a_1	a_2	a_3
a_1	1	0	1/1024
a_2	0	1	1/256
a_3	1/1024	1/256	1

Therefore, the Limited Tolerance Relation model is proposed in paper [17].

Definition 10. Given an incomplete information system $S = \langle U, A, V, f \rangle$, where $B \subseteq A$ and $P_B(x) = \{b \mid b \in B \wedge b(x) \neq *\}$, then the Limited Tolerance Relation L is defined as in (6):

$$\forall_{x,y \in U \times U}(L_B(x,y) \Leftrightarrow \forall_{b \in B}(b(x)=b(y)=*) \vee ((P_B(x) \bigcap P_B(y) \neq \varnothing) \wedge \forall_{b \in B}((b(x) \neq *) \wedge (b(y) \neq *) \rightarrow (b(x)=b(y))))) \quad (6)$$

According to the above definition, the similar object of a_3 is only a_1 in Table II, so the attribute value 1 of object a_1 should be directly chosen to fill in attribute c_4 of object a_3.

3.4. Valued Limited Tolerance Relation

In Limited Tolerance Relation model, the similarity between two objects can only be qualitatively analyzed, so its effect is not as good as quantitative analysis in Valued Tolerance Relation model. Combining the advantages of Valued Tolerance Relation model and Limited Tolerance Relation model, the Valued Limited Tolerance Relation model is proposed in this paper.

Definition 11. Given an incomplete information system $S=\langle U,A,V,f\rangle$, where $B \subseteq A$ and $P_B(x)=\{b \mid b \in B \wedge b(x) \neq *\}$, then the Valued Limited Tolerance Relation Matrix is defined as in (7):

$$M_{VL}(i,j)=\begin{cases} 1 & i=j \\ 0 & (i \neq j) \wedge (P_B(i) \bigcap P_B(j)=\varnothing) \\ \prod_{a_k \in A} P_k(i,j) & (i \neq j) \wedge (P_B(i) \bigcap P_B(j) \neq \varnothing) \end{cases} \quad (7)$$

$P_k(i,j)$ is defined as in (4).

3.5. Data Filling Algorithm Based on Valued Limited Tolerance Relation

If the original Valued Limited Tolerance Relation Matrix is M_{VL}^0, the Valued Limited Tolerance Relation Matrix after *r*-th completing analysis is M_{VL}^r. A data filling algorithm VLTA is given based on the above Valued Limited Tolerance Relation Matrix.

Input: incomplete information system $I^0=\langle U^0, AT\rangle$

Output: complete information system $I^r=\langle U^r, AT\rangle$

Step 1: Computing the original Valued Limited Tolerance Relation Matrix M_{VL}^0, MAS_i^0 ($i=1,2,...,n$) and MOS^0, let $r=0$;

Step 2:

2.1 for all $i \in MOS^r$, compute NS_i^r;

2.2 Generating I^{r+1}:

2.2.1 for all $i \notin MOS^r$, let $a_k(u_i^{r+1})=a_k(u_i^r)$, $k=1,2,...,m$;

2.2.2 for all $i \in MOS^r$, while $k \in MAS_i^r$ do:

2.2.2.1 if $|NSi^r|=1$, supposing $j \in NS_i^r$;

if $a_k(u_j^r)=*$, then let $a_k(u_i^{r+1})=*$; else let $a_k(u_i^{r+1})=a_k(u_j^r)$;

2.2.2.2 if $|NSi^r| \geq 2$

(i) if $\exists j_0, j_1, ..., j_t \in NS_i^r$, which satisfy $(a_k(u_{j_0}^r) \neq *) \vee (a_k(u_{j_1}^r) \neq *) \vee ... \vee (a_k(u_{j_t}^r) \neq *)$, let $M_{VL}^r(i,g)=\underset{1 \leq j \leq n}{MAX}(M_{VL}^r(i,j))$, then let $a_k(u_i^{r+1})=a_k(u_{j_g}^r)$;

(ii) if $\exists j_0, j_1, ..., j_t \in NS_i^r$, which satisfy $(a_k(u_{j_0}^r)=*) \wedge (a_k(u_{j_1}^r)=*) \wedge ... \wedge (a_k(u_{j_t}^r)=*)$, then let $a_k(u_i^{r+1})=*$;

2.2.2.3 if $|NS_i^r| = 0$, then let $a_k(u_i^{r+1}) = *$;

2.3 if $I^{r+1} = I^r$, then endwhile and goto Step 3;

else compute M_{VL}^{r+1}, MAS_i^{r+1} and MOS^{r+1}; $r = r+1$; goto Step 2;

Step 3: If missing data still exist in the information system, they can be filled by other algorithms such as Mean Completer and Combinatorial Completer.

Step 4: End.

4. Instances

Instance 3. Given an incomplete information system as in Table 4, where $a_1, a_2, ..., a_8$ are objects, c_1, c_2, c_3, c_4 are four attributes which values (discrete) range from 0 to 3, "*" denotes the missing attribute value.

Table 4. An Information System

A	c_1	c_2	c_3	c_4
a_1	3	2	1	0
a_2	*	2	*	0
a_3	3	3	1	*
a_4	2	2	*	0
a_5	3	*	*	3
a_6	*	2	2	*
a_7	3	2	3	3
a_8	2	*	2	*

Firstly, the incomplete information Table 4 is analyzed by using Limited Tolerance Relation. The following results can be obtained:

$I_C^L(a_1) = \{a_1, a_2\}$,

$I_C^L(a_2) = \{a_1, a_2, a_4, a_6\}$,

$I_C^L(a_3) = \{a_3, a_5\}$,

$I_C^L(a_4) = \{a_2, a_4, a_6, a_8\}$,

$I_C^L(a_5) = \{a_3, a_5, a_7\}$,

$I_C^L(a_6) = \{a_2, a_4, a_6, a_8\}$,

$I_C^L(a_7) = \{a_5, a_7\}$,

$I_C^L(a_8) = \{a_4, a_6, a_8\}$

We can get the Valued Limited Tolerance Relation Matrix as shown in Table 5.

Table 5. Valued Limited Tolerance Relation Matrix

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
a_1	1	1/16	0	0	0	0	0	0
a_2	1/16	1	0	1/64	0	1/256	0	0
a_3	0	0	1	0	1/64	0	0	0
a_4	0	1/64	0	1	0	1/64	0	1/64
a_5	0	0	1/64	0	1	0	1/16	0
a_6	0	1/256	0	1/64	0	1	0	1/256
a_7	0	0	0	0	1/16	0	1	0

a_8	0	0	0	1/16	0	1/256	0	1

The result of incomplete data filling algorithm VLTA is shown in Table 6.

Table 6. The Result of Algorithm VLTA

A	c_1	c_2	c_3	c_4
a_1	3	2	1	0
a_2	3	2	1	0
a_3	3	3	1	3
a_4	2	2	2	0
a_5	3	2	3	3
a_6	2	2	2	0
a_7	3	2	3	3
a_8	2	2	2	0

Then the incomplete information Table 4 is analyzed by using Tolerance Relation. The following results can be obtained:

$I_C^T(a_1)=\{a_1,a_2\}$,
$I_C^T(a_2)=\{a_1,a_2,a_4,a_6,a_8\}$,
$I_C^T(a_3)=\{a_3,a_5\}$,
$I_C^T(a_4)=\{a_2,a_4,a_6,a_8\}$,
$I_C^T(a_5)=\{a_3,a_5,a_6,a_7\}$,
$I_C^T(a_6)=\{a_2,a_4,a_5,a_6,a_8\}$,
$I_C^T(a_7)=\{a_5,a_7\}$,
$I_C^T(a_8)=\{a_2,a_4,a_6,a_8\}$

The result of incomplete data filling algorithm ROUSTIDA is shown in Table 7.

Table 7. The Result of Algorithm ROUSTIDA

A	c_1	c_2	c_3	c_4
a_1	3	2	1	0
a_2	*	2	*	0
a_3	3	3	1	3
a_4	2	2	2	0
a_5	3	*	*	3
a_6	*	2	2	*
a_7	3	2	3	3
a_8	2	2	2	0

By comparing the results of these two algorithms, we can find that all the missing data are filled by using algorithm VLTA based on Valued Limited Tolerance Relation. However, some attributes of objects a_2,a_5,a_6 can't be filled by using algorithm ROUSTIDA due to the conflict of attributes. On this condition, the missing data must be filled by using other approaches. It can easily lead to decision conflicts.

Instance 4. Given an incomplete information system as in Table 8, where $a_1,a_2,...,a_6$ are objects, c_1,c_2,c_3,c_4 are four attributes which values (discrete) range from 0 to 3, "*" denotes the missing attribute value.

Table 8. An Information System

A	c_1	c_2	c_3	c_4
a_1	0	*	*	*
a_2	0	1	*	*
a_3	*	1	2	*
a_4	*	*	2	3
a_5	*	2	0	2
a_6	1	2	0	2

Firstly, the incomplete information Table 8 is analyzed by using Limited Tolerance Relation. The following results can be obtained:

$I_C^L(a_1)=\{a_1,a_2\}$,

$I_C^L(a_2)=\{a_1,a_2,a_3\}$,

$I_C^L(a_3)=\{a_2,a_3,a_4\}$,

$I_C^L(a_4)=\{a_3,a_4\}$,

$I_C^L(a_5)=\{a_5,a_6\}$,

$I_C^T(a_6)=\{a_5,a_6\}$

We can get the Valued Limited Tolerance Relation Matrix as shown in Table 9.

The result of incomplete data filling algorithm VLTA is shown in Table 10.

Table 9. Valued Limited Tolerance Relation Matrix

	a_1	a_2	a_3	a_4	a_5	a_6
a_1	1	1/1024	0	0	0	0
a_2	1/1024	1	1/256	0	0	0
a_3	0	1/256	1	1/256	0	0
a_4	0	0	1/256	1	0	0
a_5	0	0	0	0	1	¼
a_6	0	0	0	0	1/4	1

Table 10. The Result of Algorithm VLTA

A	c_1	c_2	c_3	c_4
a_1	0	1	2	3
a_2	0	1	2	3
a_3	0	1	2	3
a_4	0	1	2	3
a_5	1	2	0	2
a_6	1	2	0	2

Then the incomplete information Table 8 is analyzed by using Tolerance Relation. The following results can be obtained:

$I_C^T(a_1)=\{a_1,a_2,a_3,a_4,a_5\}$,

$I_C^T(a_2)=\{a_1,a_2,a_3,a_4\}$,

$I_C^T(a_3)=\{a_1,a_2,a_3,a_4\}$,

$I_C^T(a_4)=\{a_1,a_2,a_3,a_4\}$,

$I_C^T(a_5)=\{a_1,a_5,a_6\}$,
$I_C^T(a_6)=\{a_5,a_6\}$
We can get the Valued Tolerance Relation Matrix as shown in Table 11.

Table 11. Valued Tolerance Relation Matrix

	a_1	a_2	a_3	a_4	a_5	a_6
a_1	1	1/1024	1/1024	1/1024	1/256	0
a_2	1/1024	1	1/256	1/256	0	0
a_3	1/1024	1/256	1	1/256	0	0
a_4	1/1024	1/256	1/256	1	0	0
a_5	1/256	0	0	0	1	1/4
a_6	0	0	0	0	¼	1

The result of incomplete data filling algorithm VTRIDA based on Valued Tolerance Relation is shown in Table 12.

By comparing the results of these two algorithms, the results of objects $a_2,...,a_6$ are all same. Let's consider object a_1, the result of algorithm VLTA is {0,1,2,3}. It retains the consistency among the objects in the information system. However, the result of algorithm VTRIDA is {0,2,0,2}. This result is different from all the other objects in the information system. Obviously, this is a noisy object and it can easily lead to decision conflicts.

Table 12. The Result of Algorithm VTRIDA

A	c_1	c_2	c_3	c_4
a_1	0	2	0	2
a_2	0	1	2	3
a_3	0	1	2	3
a_4	0	1	2	3
a_5	1	2	0	2
a_6	1	2	0	2

5. Analysis

5 *UCI* datasets [18] are used in order to test the feasibility and effectiveness of the algorithm VLTA. The continuous attributes of the datasets are discretized by using Semi-naive algorithm in software ROSETTA [19]. Because the original datasets don't have missing data, we tackle them with the method as follows. The random data are generated through the random data generator. The places of missing data are determined by the random data. Then the corresponding attribute values are deleted, and the incomplete information system is generated.

In the experiments, the ratios of missing data are 5% and 10%, respectively. Three algorithms ROUSTIDA, VTRIDA and VLTA are respectively used to fill in the same incomplete information system generated. If there still have missing data after filling, *Mean Completer algorithm* is adopted to fill in these remaining missing data. Correct ratio is defined as the ratio of correct samples and total missing samples. The experimental results are shown in Table 13. Among them, (1)(2)(3) denote the algorithm ROUSTIDA, VTRIDA and VLTA, respectively.

Table 13. Experimental Results

datasets	object numbers	attribute numbers	correct ratio of 5% missing data(%)			correct ratio of 10% missing data(%)		
			(1)	(2)	(3)	(1)	(2)	(3)
Car Evaluation	1728	6	83.3	85.1	85.6	83.7	83.9	84.2
Zoo	101	17	82.5	82.7	83.7	82.2	83.7	84.3
SPECT Heart(train)	267	22	78.4	79.5	82.3	77.9	78.3	80.5
Image Segmentation	2310	19	87.7	88.2	90.2	88.1	89.2	90.1
Kinship	104	12	75.1	76.2	76.2	74.1	74.8	75.8
Balance Scale	625	4	79.3	80.7	82.1	78.7	81.9	81.9

From the experimental results we can find that the filling accuracy of algorithm VLTA is higher than two other algorithms. It indicates that the VLTA makes full use of the common characteristics of datasets. It can effectively eliminate the rules caused by noisy data and has higher classification accuracy.

The experimental results of 5% and 10% missing data are shown in Figure 1 and Figure 2. Among them, 1~6 of axes X denote the dataset Car Evaluation, Zoo, SPECT Heart(train), Image Segmentation, Kinship and Balance Scale, respectively.

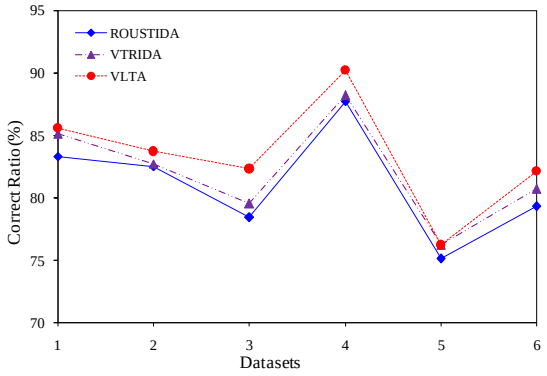


Figure 1. Experimental Results of 5% Missing Data

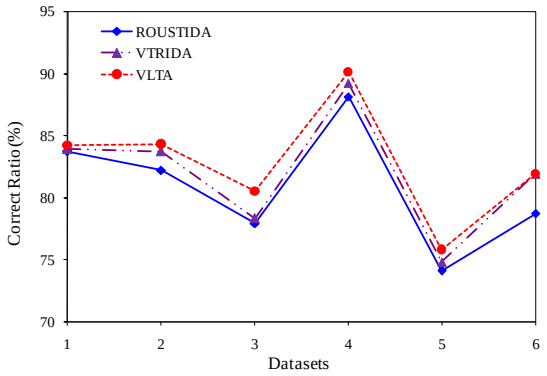


Figure 2. Experimental Results of 10% Missing Data

6. Conclusions

Tolerance Relation model, Valued Tolerance Relation model and Limited Tolerance Relation model are studied in this paper. The Valued Limited Tolerance Relation model is proposed. A data filling algorithm VLTA based on the Valued Limited Tolerance Relation model is presented. The VLTA makes full use of the advantages of Rough Set. It only needs the information which is provided by the information systems and doesn't need additional information. More missing data can be filled scientifically. The experiments show that the effect of this approach is good. It can be used as a kind of data preprocessing means in data mining. How to make further improvement on the efficiency is our future work.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. No. U1404611, U1204614, No. 61370221, and by the key project of the Education Department Henan Province under Grant No. 14B520031, in part by Program for Science & Technology Innovative Research Team in University of Henan Province under Grant No. 14IRTSTHN021, in part by the Program for Science & Technology Innovation Talents in University of Henan Province under Grant No. 14HASTIT045.

References

[1] J. W. Grzymala-Busse and M. Fu, "A comparison of several approaches to missing attribute values in data mining", In: Proc of the 2nd International Conference on Rough Sets and Current Trends in Computing. Berlin: Springer-Verlag, **(2000)**, pp. 378-385.
[2] Z. Pawlak, "Rough Sets", International Journal of Computer and Information Sciences, no. 11, **(1982)**, pp. 341-356.
[3] Z. Pawlak, J. G. Busse and R. Slowinskiet al, "Rough sets", Communications of the ACM, vol. 38, no. 11, **(1995)**, pp. 89-95.
[4] S. M. Rabiee and H. Baseri, "Prediction of the Setting Properties of Calcium Phosphate Bone Cement", Computational Intelligence and Neuroscience, Article ID 809235, **(2012)**, pp. 8.
[5] Z. Wei, L. X. Feng and W. Z. Fu. "An incomplete data analysis approach based on rough set theory", Pattern Recognition and Artificial Intelligence, vol. 16, no. 2, **(2003)**, pp. 158-163.
[6] Z. Z. Hua and L. W. Qi, "An Improved Algorithm Based on the Incomplete Data of the Rough Set Theory", Computer Engineering & Science, vol. 24, no. 4, **(2002)**, pp. 41-42, 67.
[7] F. Yuan, "A Method of Recruiting Default in Data Based on the Decision", Journal of Kunming University of Science and Technology (Science and Technology), vol. 28, no. 6, **(2003)**, pp. 157-160.
[8] T. S. Xin, W. X. Ping and W. H. Xia, "Improved method for data reinforcement based on ROUSTIDA", Journal of Naval University of Engineering, vol. 23, no. 5, **(2011)**, pp. 11-15.
[9] M. Kryszkiewicz, "Rough set approach to incomplete information system", Information Sciences, vol. 112, **(1998)**, pp. 39-49.
[10] M. Kryszkiewicz, "Properties of incomplete information systems in the framework of rough sets", In: L Polkowski, A Skowron eds. Rough Sets in Data Mining and Knowledge Discovery. Berlin: Springer-Verlag, **(1998)**, pp.422-450.
[11] Z. X. Fei and Z. L. Xia, "An Incomplete Data Analysis Method Based on the Valued Tolerance Relation", Journal of Chongqing Institute of Technology, vol. 19, no. 5, **(2005)**, pp. 23-25.
[12] Z. Y. Ming, L. X. Qin and C. Z. Xin, "Completion of Incomplete Intelligence Information system Based on Rough Set", Military Operations Research and Systems Engineering, , vol. 21, no. 2, **(2007)**, pp. 50-53.
[13] D. C. Rong and L. L. Shu, "Completing data algorithm based on similarity relation vector", Application Research of Computers, vol. 30, no. 2, **(2013)**, pp. 383-385.
[14] J. Stefanowski and A .Tsoukias, "On the extension of rough sets under incomplete information", In: N Zhong, A Skowron, S Ohsuga eds. Proc of the 7th Int' l Workshop on New Directions in Rough Sets, Data Mining, and Granular-Soft Computing. Berlin: Springer-Verlag, **(1999)**, pp.73-81.
[15] Skowron A. and Rauszer C., "The Discernibility Matrixes and Function in Information System", Intelligent Decision Support-Handbook of Applications and Advances of the Rough Set Theory, **(1991)**, pp.331-362.

[16] R. Zdunek and A. Cichocki, "Fast Nonnegative Matrix Factorization Algorithms Using Projected Gradient Approaches for Large-Scale Problems", Computational Intelligence and Neuroscience, Article ID 939567, **(2008)**, pp. 13.
[17] W. G. Yin, "Extension of Rough Set under Incomplete Information Systems", Journal of Computer Research and Development, vol. 39, no. 10, **(2002)**, pp. 1238-1243.
[18] Merz, C. J. and Murphy P., "UCI repository of machine learning database", http://archive.ics.uc*i.e*du/ml/.
[19] Rosetta, "A Rough Set Toolkit for Analyzing Data", http:www.idi.ntnu.no/~aleks/rosetta/.

Authors

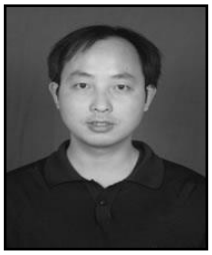
Xiuling Bai, She was born in Henan Province, PRC in Aug.1974. Xiuling Bai studied in Henan University of Science and Technology (Luoyang, Henan Province, PRC) from Sept.2000 to Jun.2003, majored in computer application and earned a Master of Engineering Degree in three year's time. She works as an Associate Professor in Henan University of Science and Technology. Her research interests include Artificial Intelligence, Data Mining and Rough Set.

Mingchuan Zhang, He was born in Henan Province, PRC in May 1977. Mingchuan Zhang studied in Beijing University of Posts and Telecommunications (Beijing, PRC) from September 2011 to July 2014, majored in Communication and information system and earned a Doctor of Engineering Degree in three year's time. He works as an Associate Professor in Henan University of Science and Technology. His research interests include bio-inspired networks, Internet of Things, future Internet and computer security.

Qingtao Wu, He was born in Jiangsu Province, PRC in Mar 1975. Qingtao Wu studied in East China University of Science and Technology (Shanghai, PRC) from Mar 2003 to Mar 2006, majored in computer application and earned a Doctor of Engineering Degree in three year's time. He works as a Professor in Henan University of Science and Technology. His research interests include component technology, computer security and future Internet security.

Ruijuan Zheng, She was born in Henan Province, PRC in Mar 1980. Ruijuan Zheng studied in Harbin Engineering University Technology (Harbin, PRC) from Mar 2005 to Mar 2008, majored in computer application and earned a Doctor of Engineering Degree in three year's time. She works as an Associate Professor in Henan University of Science and Technology from Mar 2008 to now. Her research interests include bio-inspired networks, Internet of Things, and computer security.

Haixia Zhao, She was born in Henan Province, PRC in July 1976. Haixia Zhao studied in National University of Defense Technology (Changsha, Hunan Province, PRC) from September 2001 to July 2006, majored in computer application and earned a Master of Engineering Degree in three year's time. She works as an Associate Professor in Henan University of Science and Technology. Her research interests include sensor networks, and computer security.



Wangyang Wei, He was born in Henan Province, PRC in Mar 1979. Wangyang Wei is a Doctor student of Beijing University of Posts and Telecommunications (Beijing, PRC). He works as a Lecturer in Henan University of Science and Technology. His research interests include Internet of Things, and computer security.