

A Method for Filling up the Missed Data in Information Table

Gao Xuedong

School of Management, University of Science and Technology Beijing, P.R. China, 100083
Phone: 86-010-62333706, 86-010-62333582, Email: gaoxuedong@manage.ustb.edu.cn

E Xu

School of Management, University of Science and Technology Beijing, P.R. China, 100083
Phone: 86-010-62333706, 86-010-62333582, Email: exu21@163.com

Li Tieke

School of Management, University of Science and Technology Beijing, P.R. China, 100083
Phone: 86-010-62333733, 86-010-62333582, Email: tieke@manage.ustb.edu.cn

Zhang Qun

School of Management, University of Science and Technology Beijing, P.R. China, 100083
Phone: 86-010-62332744, 86-010-62333582, Email: zq@manage.ustb.edu.cn

ABSTRACT

Almost algorithms based on the rough sets, such as mean value method, maximum frequency method, mode method are weak in supporting the hidden rules in the information table. By the breaking point sets in the information system, a new method for packing the missed attribute value is provided in the paper. The method is more efficient for indicating the decision rules.

Key words: Rough sets, Information table, Classing; Decision rule, Breaking point

INTRODUCTION

The on-going information revolution is generating volumes of data, from sources as diverse as banking transactions, scientific explorations, telecommunication networks, space science, medical systems and so on. Indeed, there is an imperative on the intelligent analysis of such large volumes of data so as to derive intrinsic knowledge (Krysiakiewicz, 1998), which can impact to optimize decision support, business competitiveness and other services-oriented portfolios.

Rough sets theory (Wu et al., 2003; Michalski, 1986; Pawlak, 1982) has been proposed by Professor Pawlak for knowledge discovery in databases and experimental data sets. It is based on the concept of an upper and a lower approximation of rough sets, indiscernible matrix and so on. Rough sets theory (Kohavi and Frasca, 1994) is an efficient mathematical tool to deal with the uncertain and incomplete data, which don't need some transcendental knowledge or some accessional information but just the data itself. For now, it has been applied in many fields such as machine learning, artificial intelligence and so on; especially it has been a very efficient method in data mining, which frequently appears in clustering, classification algorithms.

In rough sets theory, data exists in the information table. But because of various reasons such as wrong operation, off power and so on, there are always some missed data in the information table, we must pack them so as to analyze them.

For now, there have been many algorithms (Liu and Setiono, 1995; Pawlak, 1991; Wang, 2000) for filling up the missed data in the information table, which can be divided into two categories in terms of thinking or not thinking about rough sets, such as mean value method that takes the whole data average on some certain attribute as the missed data, maximum frequency method that takes the most frequent data on some certain attribute as the missed data, mode method that takes the middle value on some certain attribute as the missed data, and so on. But those methods can not support the hidden rules in the information table very well because they think little or at least don't think about rough sets.

A new approach for filling up the missed attribute value is presented in the paper based on the breaking point sets in the information system, which is presented after studying the correlation of conditional and decision attributes. Some experiments show that the method is more efficient for indicating the decision rules.

CHIEF CONCEPTS OF ROUGH SETS THEORY

Rough sets theory includes some significant concepts (Wang, 2000; Pawlak, 1998) such as an upper approximation, a lower approximation, etc.

In rough set theory, annotated data is represented as an information system. An information system can be represented as

$$S = (U, A, V, f),$$

Where U is the universe, a finite set of N objects $U = \{x_1, x_2, \dots, x_n\}$, A is a finite set of attributes, which are divided into disjoint sets, i.e. $A = C \cup D$, where C is the set of condition attributes and D is the set of decision attribute. $V = \bigcup_{q \in A} V_q$ (where V_q is a domain of the attribute q), $f : U \times A \rightarrow V$ is the total decision function (called information function), such that $f(x, q) \in V_q$ for every $q \in A, x \in U$. A subset of attributes $Q \subseteq A$ defines an equivalence relation (called an indiscernible relation) on U, $IND(Q) = \{(x, y) \in U : \forall a \in Q, f(x, a) = f(y, a)\}$.

For a given information system $S = (U, A, V, f)$, a subset of attributes $Q \subseteq A$ determines the approximation space $U | IND(Q)$ in S. For given $Q \subseteq A$ and $X \subseteq U$ (a concept X), the Q lower approximation $\underline{Q}X$ of the set X and the Q upper approximation $\overline{Q}X$ of the set X are defined as follows:

$$\underline{Q}X = \{x \in U : [x]_A \subseteq X\} \tag{1}$$

$$\overline{Q}X = \{x \in U : [x]_A \cap X \neq \emptyset\} \tag{2}$$

And meanwhile, we can define the Q-positive region $POS_Q(D)$ in the relation $IND(D)$ as

$$POS_Q^U(D) = \bigcup \{ \underline{Q}X : X \in IND(D) \} \tag{3}$$

The positive region $POS_Q(D)$ contains all the objects in U that can be classified without an error into distinct classes defined by $IND(D)$. So we can define the rough set boundary as $U' = U - POS_Q^U(D)$. One can form a positive region for any two subsets of attributes $B, E \in A$ in the information system S. Since the subset of attributes $B \in A$ defines the indiscernible relation $IND(B)$, it consequently defines the classification $B^* (B^* (U | IND(B)))$ with respect to the subset E. The E-positive region of B is defined as

$$POS_E(B) = \bigcup_{X \in B^*} \underline{E}X . \tag{4}$$

The E-positive region of B contains all the objects that, by using attributes E, can be classified to one of distinct classes of the classification B^* .

The cardinality of the E-positive region of B can be used to define a measure $\gamma_E(B)$ of dependency of the set of attributes B on E:

$$\gamma_E(B) = \frac{card(POS_E(B))}{card(U)} \tag{5}$$

Rough sets define a measure of significance (coefficient of significance) of the attribute $a \in A$ from the set E with respect to the subset B :

$$sig(a) = \gamma_A(B) - \gamma_{A \setminus \{a\}}(B). \tag{6}$$

PMMDVBBP ALGORITHM

PMMDVBBP (Packing Method for Missed Data Value Based on Breaking Points) algorithm is based on the breaking points in the information table with respect to the correlation of condition and decision attributes, so it can padded the missed data better.

Relevant Definition and Theorem

Given an information system $S=(U,A,V,f)$, $A = \{a_i \mid i = 1, \dots, m\}$ is the attribute set, assumed $X_i \in U$, then it's missed attributes set MAS_i , indiscernible set NS_i and the missed object set of the information system MOS are defined as follows:

$$MAS_i = \{a_k \mid a_k(x_i) = *, k = 1, \dots, m\} \tag{7}$$

$$MOS = \{i \mid MAS_i \neq \emptyset, i = 1, \dots, n\} \tag{8}$$

$$NS_i = \{j \mid M(i, j) = \emptyset, i \neq j, j = 1, \dots, n\} \tag{9}$$

From the information system table, we can proof a theorem as follows:

Theorem: Given an information system $S = \langle U, A, V, F \rangle$, where $U = U^0 \cup U'$, U^0 is the total sample set with complete attribute values, U' is the sample set that we only know the partial values. $A = C^0 \cup C' \cup D$, C^0 is the significant attribute set, C' is the redundant attribute set, D is the decision attribute set. If $\forall a \in U', \forall b \in U^0, \forall c \in C', c(a) = c(b)$, then conclude that the information system's certainty is stable.

Proof: We can assume any classification of the system, $E_i \in U \mid IND(C), (i = 1, 2, \dots, m)$, m is the number of the classification divided by the condition attribute set C , $\{X_1, X_2, \dots, X_n\} = U \mid IND(D)$, then for some certain classification $E \in U \mid IND(C)$, it's certainty to the decision attribute class is as follows:

$$\mu_{\max}(E) = \max(\{|E \cap X_i| / |E| : X_i \in U \mid IND(D)\}) \tag{10}$$

We can induce the information system certainty as follows:

$$\mu_{\max}(S) = \sum_{i=1}^m \frac{|E_i|}{|U|} \cdot \mu_{\max}(E_i) \tag{11}$$

Based on the formula (11), we can discuss the above theorem from two angles:

- If C' has only one element c , then we can regard the formula $E_i \in U \mid IND(C^0 \cup \{c\}), (i = 1, 2, \dots, m')$ as the classes determined by the condition attribute set $C = C^0 \cup \{c\}$. Since c is the redundant attribute, we consequently induce the equation $U \mid IND(C^0) = U \mid IND(C^0 \cup \{c\})$, namely adding redundant attribute can't affect the classes in the information system S , i.e. $E = E'$, therefore, we can induce result that if $\forall E, E \in U \mid IND(C^0)$, then there must be $\exists E \in U \mid IND(C^0 \cup \{c\})$, we can conclude the formula $\mu_{\max}(E) = \mu_{\max}(E')$; so obtain the conclusion that $\mu_{\max}(S)$ in the information table is not changed.
- In the same way, if $C' = \{C'_1, C'_2, \dots, C'_m\}$ is the redundant attribute set, then the $\mu_{\max}(S)$ in the information table will not changed

PMMDVBBP Algorithm Description

Based on the chief concepts of rough sets theory, especially the above formula (8), formula (9), formula (10) and formula (11), the paper puts forward an algorithm for filling up the missed data in the information system S ,

PMMDVBBP algorithm. The main steps of PMMDVBBP algorithm can be described as follows, given the initial information system table $S^0 = \langle U^0, A^0, V^0, f^0 \rangle$.

- Step 1: Calculate the relevant coefficient among various attributes, construct the matrix in $U \setminus MOS$.
- Step 2: Cluster the condition attributes in the relevant coefficient matrix, then select one representative attribute from every cluster to form a new condition attribute set.
- Step 3: Calculate the condition attribute importance in $U \setminus MOS$, sort them by the importance, and then place MOS at the end of the information table to form a new information table $S^i = \langle U^i, A^i, V^i, f^i \rangle$.
- Step4: Assumed that P^i is the breaking point set of anyone attribute $A^i, A^i \in U \setminus MOS$, the breaking point is $P_j^i, j=1,2,\dots,n-1$ its nearest neighbors are x_j^i, x_{j+1}^i , furthermore, $x_j^i < P_j^i < x_{j+1}^i$, we can deal with A^i as follows:
 FOR (i=1; n; i++)
 {if $\alpha = x_j^i; x_j^i = x_{j+1}^i$, the information table has no conflicts, then $P^i = P^i \setminus \{P_j^i\}$;
 else $x_j^i = \alpha; x_{j+1}^i = x_{j+1}^i$.}
- Step 5: Classify the samples in the information table, sort the classes by their cardinality, and then operate them as Step 6.
- Step 6: We can assume any condition attributes set $B^i \subseteq B \subseteq A, \forall x_i \in MOS$, and operate them as follows in one class:
 (1) If $a_k \in B^i$ and $r_B(F) = r_{B \setminus B^i}(F)$, then let $a_k(x_i)$ is equal to any number.
 (2) Else operate as follows:
 ① If $\forall x_i \in MOS$ calculate $NS(x_i)$ in the attribute set $B \setminus \{B^i\}$
 ② If $NS(x_i) = y$ then let $a_k(x_i) = a_k(y)$
 ③ If $NS(x_i) = \{y_1, y_2, \dots, y_m\}$ then let $y = y_i$ the cardinality of y_i is max in the $NS(x_i)$ and $a_k(x_i) = a_k(y)$
 (3) Operate the information table as above until $x_i \in MOS$ is the final attribute.
- Step 7: Select the second class in the decision table and operate it as Step 6.
- Step 8: Operate the decision table as Step 4 until the final class, then we can export a new information table $S = \langle U, A, V, f \rangle$.

The above steps are the contents of PMMDVBBP algorithm proposed by the paper.

DEMONSTRATION

In order to explain the thinking and contents of PMMDVBBP algorithm, we give a demonstration.
 Example: Given an initial information system $S = \langle U, A, V, f \rangle$, where U is the universe, a finite set of N objects $U = \{x_1, x_2, \dots, x_n\}$, A is a finite set of attributes, which are divided into disjoint sets, i.e. $A = C \cup D$, where C is the set of condition attributes and D is the set of decision attribute. The contents of it in details are showed as Table 1.

U	a	b	c	d
1	0.9	2	1	1
2	1.1	0.8	1	0
3	1.3	3	2	0
4	1.4	1	1	1
5	1.4	2	1	1
6	1.2	1	1	1
7	1.8	3	2	1
8	4	3	2	1
9	*	3	2	1
10	1.3	*	1	0

11	2	1	*	0
----	---	---	---	---

Table 1: The initial information table.

According to some important concepts of rough set and the chief steps of PMMDVBBP algorithm, we can do as follows:

$$\begin{aligned}
 U \mid IND(a,b,c) &= \{\{1\}, \{2\}, \dots, \{8\}\} \\
 U \mid IND(a) &= \{\{1\}, \{2\}, \{3\}, \{4,5\}, \{7\}, \{8\}\} \\
 U \mid IND(b) &= \{\{1,5\}, \{2\}, \{3,7,8\}, \{4,6\}\} \\
 U \mid IND(c) &= \{\{1,2,4,5,6\}, \{3,7,8\}\} \\
 r_c(D) &= card(POS_c(D)) / card(U) = 8/8 = 1 \\
 r_{c \setminus \{a\}}(D) &= card(POS_{c \setminus \{a\}}(D)) / card(U) = 3/8 \\
 r_{c \setminus \{b\}}(D) &= card(POS_{c \setminus \{b\}}(D)) / card(U) = 5/8 \\
 r_{c \setminus \{c\}}(D) &= card(POS_{c \setminus \{c\}}(D)) / card(U) = 8/8
 \end{aligned}$$

So, the importance of attribute a is equal to $r_c(D) - r_{c \setminus \{a\}}(D) = 1 - 0.375 = 0.625$,

the importance of attribute b is equal to $r_c(D) - r_{c \setminus \{b\}}(D) = 1 - 0.625 = 0.375$,

the importance of attribute c is equal to $r_c(D) - r_{c \setminus \{c\}}(D) = 1 - 1 = 0$ obviously, the attribute c is redundant. Following the other steps, we can obtain the Table 2. and Table 3 from Table 1.

U	a	b	c	d
1	0.9	2	3	1
4	1.4	1	1	1
6	1.2	1	1	1
7	1.8	3	3	1
8	4	3	3	1
9	*	3	2	1
2	1.1	1	1	0
3	1.3	3	3	0
5	1.4	2	3	0
10	1.3	*	1	0
11	2	1	*	0

Table 2: The middle table.

U	a	b	c	d
1	0.9	2	3	1
4	1.4	1	1	1
6	1.2	1	1	1
7	1.8	3	3	1
8	4	3	3	1
9	4	3	2	1
2	1.1	1	1	0
3	1.3	3	3	0
5	1.4	2	3	0
10	1.3	3	1	0
11	2	1	1	0

Table 3: The final table that has been filled up.

For test the performance of the algorithm above, an experiment has been done about the famous Iris classification problem in the UCI machine learning database. There are only instances with continuous attributes in Iris database. Every instance has four continuous attributes, which are petal-length, petal-width, sepal-length, sepal-width. And the sum number of the instances, which belong to three classes, is 150. The experiment result is as Table 4. PMMDVBBP algorithm is proved to be reliable and efficacious from Table 4.

Number(missed data) (missed data)	10	15	30	40
Number(padded correctly) (padded correctly)	8	12	23	32

Table 4: Experimental result.

CONCLUSION

This algorithm is reliable, efficient and simple. It is proposed on the correlation of condition attributes and decision attributes, especially on the importance of the breaking points in the information system. So it avoids conflicts in the information system that general algorithms, such as the mean method, the maximum method, the mod method and so on, made in the information system and the data that it filled up can indicate the hidden knowledge better.

REFERENCES

- Kohavi R, Frasca B. (1994). Useful Feature Subsets and Rough Set Reducts. *The 3th International Workshop on Rough Sets and Soft Computing*.
- Krysiakiewicz M. (1998). Rough Set Approach to Incomplete Information System. *Information Sciences*, 112, 39~49.
- Liu H., Setiono R. (1995). Feature Selection and Discretization of Numeric Attributes. *Proceedings of the 7th International Conference on Tools with Artificial Intelligence*, Washington D.C.
- Michalski R S. (1986). Multi-Purpose Incremental Learning System AQ15 and Its Testing Application to Three Medical Domains. *Proceedings of the 5th AAAI*, CA.
- Pawlak Z. (1982). Rough Set. *International Journal of Computer and Information Science*. 1, 341-356.
- Pawlak Z. (1991). *Rough Sets-Theoretical Aspects of Reasoning about Data*. Kluwer: Kluwer Academic Publisher.
- Pawlak Z. (1998). Rough Set Theory and Its Application to Data Analysis. *Cybernetics and Systems*, 29 (9), 661-668.
- Wang Guo-yin. (2000) *Rough Set Theory and Knowledge Acquisition (in China)*. Xi'an, China: Xi'an Communication University Publisher.
- Wu Sen, Gao Xue-dong, M. Bastian. (2003). *Data House and Data Mining*. Beijing, China, Metallurgy Industrial Publisher.