Proceedings of the International Conference of Machine Vision and Machine Learning
Prague, Czech Republic, August 14-15, 2014
Paper No. 136

Rearrangement of Attributes in Information Table and its Application for Missing Data Imputation

Gongzhu Hu
Department of Computer Science, Central Michigan University
Mount Pleasant, Michigan, USA
hu1g@cmich.edu

Feng Gao
Science School, Qingdao Technological University
Qingdao, China
gaofeng99@sina.com

Abstract- Missing data is a major problem for most data analysis tasks. Many methods have been developed to address this problem, including imputation of the missing values. Imputation methods based on rough set theory have been proposed in the literature and shown to be effective. In rough set theory, data is usually stored in an information table with attributes divided into condition attributes and decision attribute. Due to the uncertainty in the data, The data set is represented by formal approximations and "condition $\rightarrow$ decision" rules can be deducted from the approximations. In this paper, we propose an approach to the missing value imputation problem by rearranging the attributes such that the attribute with missing values becomes the decision attribute so that decision rules deducted can be used to determine the missing values. For this purpose, we introduce the notion of optimal logic attribute and optimal attribute logical flow based on the roughness of rearrangements to explore the logical causal relations between attributes. Such relations can be used for missing value imputation illustrated with a few simple examples.

Keywords: Rough set, Rearrangement of attributes, Roughness of rearrangement, Optimal attribute logical flow, Missing data imputation.

1. Introduction

In rough set theory (Pawlak 1982), (Pawlak, Grzymala-Busse, Slowinski & Ziarko 1995), (Pawlak & Skowron 2007), the information of a real world application is normally expressed as an *information table* that represents the data for the application. A simple example is given in Table 1 that shows the possible results of a physician's diagnosis of six patients.

In this table, $e_1, e_2, e_3, e_4, e_5, e_6$ are called *cases* (also called objects, records, or observations). The cases are associated with *attributes*. The attributes are divided into two categories: *condition attributes* and *decision attributes*.

The basic idea of using rough set for data analysis is for make predictions based on the available data as decision rules, in the form of $condition \rightarrow decision$, that are derived from the rough sets in the data. So, *can we make decisions on the missing values (thus the missing values are imputed) rather than on the original decision attributes*? We proposed a new method to answer this question in this paper. The main idea, *that we are not aware of anyone proposed before*, is to treat an attribute (column in an information table)

Table 1. An information table

Case	Condition			Decision
	headache	*muscle_pain*	*temperature*	*flu*
e1	yes	yes	normal	no
e2	yes	yes	high	yes
e3	yes	yes	very high	yes
e4	no	yes	normal	no
e5	no	no	high	no
e6	no	yes	very high	yes

with missing value as the decision target, and the original decision target is considered a regular condition attribute. The columns of the information table are permuted (rearranged) so that each attribute column with missing values has a chance to be treated as the decision target.

The main contributions of this paper are:

1. New concepts: *roughness of rearrangement* based on the upper and lower approximations of rough set, *optimal logical attribute*, and *optimal logical attribute flow*.

2. New method: We propose a method to support decision making in missing data imputation using the attribute rearrangement based on these concepts.

2. Attribute Rearrangment

In this section, we shall present a new method using rough set that can be used for missing value imputation. We assume that the readers are familiar with the basic concepts of rough set, including indiscernibility relation, reduct, definable set and rough set, lower and upper approximations, and decision rules. Details of these concepts and definitions can be found in the literature, such as (Pawlak et al. 1995), (Pawlak & Skowron 2007).

2.1. Attribute Rearrangement

For a given information table $T = (U, A, V, f)$ where $A = C \cup D$ with the set of condition attributes $C = \{a_1, \cdots, a_k\}$ and decision attribute $D = \{d\}$, we can create a new information table $\mathscr{T} = (U, A', V, f)$ where A' is a *rearrangement* of A: $A' = C' \cup D'$, where $C' = (C - \{a_i\}) \cup \{d\}$ and $d' = \{a_i\}$. That is, the original decision attribute is swapped with a condition attribute a_i so that a_i becomes the new decision attribute.

For example, by swapping the decision attribute flu with the each of the condition attributes in Table 1, we obtain a new information table shown in Table 2(a), 2(b), and 2(c), respectively.

3. Roughness of Rearrangement and Optimal Logic

In this section, we will introduce the concept of roughness of rearrangement and associated properties that lays a foundation for a method that can be used for missing value imputation. In particular, we introduce the concept of optimal logical attribute and optimal logical attribute flow.

Definition 1 (**roughness**). The *roughness* of rearragement $\mathscr{T}$ on concept Y is

$$\beta(\mathscr{T}_Y) = \frac{|\overline{A}(Y) - \underline{A}(Y)|}{|\overline{A}(Y)|} \tag{1}$$

where $|x|$ is the cardinality of the set x.

Table 2. Rearrangments of information table

(a) headache as decision

Case	Condition			Decision
	flu	*muscle_pain*	*temperature*	*headache*
e1	no	yes	normal	yes
e2	yes	yes	high	yes
e3	yes	yes	very high	yes
e4	no	yes	normal	no
e5	no	no	high	no
e6	yes	yes	very high	no

(b) temperature as decision

Case	Condition			Decision
	headache	*muscle_pain*	*flu*	*temperature*
e1	yes	yes	no	normal
e2	yes	yes	yes	high
e3	yes	yes	yes	very high
e4	no	yes	no	normal
e5	no	no	no	high
e6	no	yes	yes	very high

(c) muscle_pain as decision

Case	Condition			Decision
	headache	*flu*	*temperature*	*muscle_pain*
e1	yes	no	normal	yes
e2	yes	yes	high	yes
e3	yes	yes	very high	yes
e4	no	no	normal	yes
e5	no	no	high	no
e6	no	yes	very high	yes

Since $|\underline{A}(Y)| \leq |\overline{A}(Y)|$, it is clear that $0 \leq \beta(\mathscr{T}_Y) \leq 1$. From the definitions of upper and lower approximations, the roughness $\beta(\mathscr{T}_Y)$ is actually a measure of the *certainty* of the logical relationship $C \to D$ in the rearrangement $\mathscr{T}$. When $\beta(\mathscr{T}_Y)$ is close to 1, the certainty is small, whereas when $\beta(\mathscr{T}_Y)$ is close to 0, the certainty is large.

For the information table and its various rearrangements in Table 1–2(c), we can calculate the roughness of some concepts as shown in Table 3.

Table 3. Calculation of roughness of rearrangements $\mathscr{T}^{(i)}$

$\mathscr{T}^{(i)}$	Concept Y	$\overline{A}(Y)$ $\underline{A}(Y)$	Roughness $\beta(\mathscr{T}_Y^{(i)})$
$\mathscr{T}^{(1)}$	*flu* = yes	$\{e2,e3,e6\}$ $\{e2,e3,e6\}$	(3-3) / 3 = 0
$\mathscr{T}^{(2(a))}$	*headache* = yes	$\{e1,e2,e3,e4,e6\}$ $\{e2\}$	(5-4) / 5 = 0.8
$\mathscr{T}^{(2(b))}$	*temperature*=very high	$\{e2,e3,e6\}$ $\{e6\}$	(3-1) / 3 = 0.67
$\mathscr{T}^{(2(b))}$	*temperature* = normal	$\{e1,e4\}$ $\{e1,e4\}$	(2-2) / 2 = 0
$\mathscr{T}^{(2(c))}$	*muscle_pain* = yes	$\{e2,e3,e6\}$ $\{e2,e3,e6\}$	(3-3) / 3 = 0

Roughness of a rearrangement $\mathscr{T}_Y$ on concept Y can be considered as an indicator of the logical relation between the condition attributes and the decision attribute. The lower the value of $\beta(\mathscr{T}_Y)$, the higher certainty of the logical relation. When roughness is 0, the logical relation $C \to D$ is completely certain.

Definition 2 (**optimal logic concept**)**.** Let $\mathscr{T}$ be a rearrangement of an information table with k concepts $Y_1, \cdots, Y_k$ defined by the decision attribute. Y_i is called the *optimal logic concept* if the roughness $\beta(\mathscr{T}_{Y_i})$ is the smallest:

$$\beta(\mathscr{T}_{Y_i}) = \min_{1 \leq j \leq k} (\beta(\mathscr{T}_{Y_j}))$$

For example, in the rearrangements $\mathscr{T}^{(i)}$ given in Tables 1–2(c) the optimal logic concepts are:

Rearrangement	Optimal Logic Concept Y
$\mathscr{T}^{(1)}$	*flu* = yes; *flu* = no
$\mathscr{T}^{(2(a))}$	*headache* = yes; *headache* = no
$\mathscr{T}^{(2(b))}$	*temperature* = normal
$\mathscr{T}^{(2(c))}$	*muscle_pain* = yes; *muscle_pain*=no

An optimal logic concept represents a most certain logical relation $C \to D$ in an rearrangement. For example, in the rearrangement $\mathscr{T}^{2(b)}$, when the values of *flu*, *headache*, and *muscle_pain* are given, we can conclude about whether the *temperature* in normal with the highest certainty, but the conclusion about the *temperature* is high or very high is less certain.

Definition 3 (**optimal logical flow of attributes**)**.** An ordered list of attributes $a_{k_1} \to a_{k_2} \to \cdots \to a_{k_n}, 1 \leq k_j \leq n$ is an *optimal logical flow* of information table T for $\beta(\mathscr{T}^{(k_1)}_{Y_{x_1}}) \geq \beta(\mathscr{T}^{(k_2)}_{Y_{x_2}}) \geq \cdots \geq \beta(\mathscr{T}^{(k_n)}_{Y_{x_n}})$.

The procedure to calculate an optimal flow is given in Algorithm 1.

Algorithm 1: Optimal logical flow

Input: $T = (U, A, V, f)$ — an information table with $A = \{a_1, \cdots, a_n\}$
Input: $S = (a_1 = v_1, \cdots, a_n = v_n)$ — a selection of attribute values
Output: Optimal logical flow of T for S

```
1  begin
2      foreach a_i ∈ A do
3          Create a rearrangement T^(i) with a_i as the decision attribute.
4          Let Y_{a_i=v_i} be the selected concept.
5          Calculate roughness β(T^(i)_{Y_{a_i=v_i}}).
6      end
7      sort β(T^(i)_{Y_{a_i=v_i}}), i = 1, ..., n in descending order.
8      Let the attributes in the sorted list be a_{k_1}, ..., a_{k_n}.
9      Create a list L with attributes a_{k_1}, ..., a_{k_n}, in that order.
10     return L.
11 end
```

Optimal logical flow indicates the logical relationships among the attributes in an information table under a group of selected concepts. The last attribute in the ordered list, a_{k_n}, which yields the smallest roughness value, is the optimal logic attribute. The logical relationship $(A - \{a_{k_n}\}) \to \{a_{k_n}\}$ has the best fit with the observed data.

Let's consider the information table in Table 1 as an example. For the attributes (*headache*, *muscle_pain*, *temperature*, *flu*), we take (yes, yes, very high, yes) as the selected concepts. For each of the attributes as the decision attribute (and hence the rearrangements in Tables 2(a)–2(c)), the roughness values are (see Table 3):

$$\begin{aligned} \beta(\mathscr{T}^{(1)}_{flu=yes}) &= 0 \\ \beta(\mathscr{T}^{(2(a))}_{headache=yes}) &= 0.8 \\ \beta(\mathscr{T}^{(2(b))}_{temperature=very\,high}) &= 0.67 \\ \beta(\mathscr{T}^{(2(c))}_{muscle\_pain=yes}) &= 0 \end{aligned}$$

The ordering of these values is $0.8 > 0.67 > 0 \geq 0$ and their corresponding concepts are (*headache*=yes, temperature=very high, *muscle_pain*=yes, *flu*=yes). Hence, the attribute flow $headache \rightarrow temperature \rightarrow muscle\_pain \rightarrow flu$ is an optimal flow. This means that these attributes with (yes, yes, very high, yes) values reflects a logical implication relationship among the attributes based on the observed data. Here either *muscle_pain*=yes or *flu*=yes can be the optimal logical attribute.

4. Missing Value Imputation with Rearrangement of Attributes

Missing value is a persistent problem for almost all data analysis tasks in the real world. Many approaches were proposed in the literature to deal with missing values (Donders, van der Heijden, Stijnen & Moons 2006), (Rubin 2009), (Schafer 2010). Most of these approaches did not consider the logical relationships between attributes. In this section, we apply the attribute rearrangement idea to the missing data imputation problem. The basic idea is to create a rearrangement of the original information table such that the attribute with to-be-imputed missing data becomes the decision attribute, and then find the logical relationship between this attribute and other attributes. If the relationship is strong, we can use the decision rules derived from the rough set theory to determine the value of the missing items; on the other hand, if the relationship is weak, we then impute the missing items using traditional statistic approach such as more frequent value replacement. This process is outlined in Algorithm 2.

Algorithm 2: Imputation with rearrangement

Input: $T = (U, A, V, f)$ — an information table with $A = \{a_1, \cdots, a_n\}$
Input: a_m — the attribute with to-be-imputed missing data
Input: v — a value of a_m
Input: b — threshold of roughness measure
Output: T' — information table of T with missing data under a_m imputed

```
1  begin
2      Create a rearrangement T' from T with a_m as the decision attribute.
3      Let Y_{a_m=v} be the selected concept.
4      Calculate roughness β(T'_{Y_{a_m=v}}).
5      if β(T'_Y) ≤ b or a_m is optimal logical attribute then
6          Derive decision rules for T'.
7          Assign values for missing items on a_m based on decision rules
8      else
9          Assign values for missing items on a_m using most frequent value.
10     return T'.
11 end
```

We now illustrate this proposed imputation approach with two examples based on the information table in Table 4.

Example 1. In this example, the value on the *headache* attribute of $e8$ is missing as shown given in Table 5(a) with $*$ representing the missing value.

By rearranging the attributes to make *headache* (that has missing value) the decision attribute, the rearrangement T' is shown in Table 5(b) with cases of complete data (i.e. case $e8$ with missing value is excluded).

For the selected attribute group (*flu*=*yes*, *temperature*=*high*, *headache*=*yes*), the roughness measure is $\beta(T'_{headach=yes}) = 1$ indicating that the logical relationship between (*flu*, *temperature*) and *headach* is weak.

Table 4. Information table with minimal reduct attributes

Case	Condition		Decision
	headache	*temperature*	*flu*
e1	yes	normal	no
e2	yes	high	yes
e3	yes	very high	yes
e4	no	normal	no
e5	no	high	no
e6	no	very high	yes
e7	no	high	yes
e8	no	very high	no

Table 5. Imputation for missing value on *headache*

(a) A value on *headache* is missing

Case	Condition		Decision
	headache	*temperature*	*flu*
e1	yes	normal	no
e2	yes	high	yes
e3	yes	very high	yes
e4	no	normal	no
e5	no	high	no
e6	no	very high	yes
e7	no	high	yes
e8	$*$	very high	no

(b) *headache* as decision attribute

Case	Condition		Decision
	flu	*temperature*	*headache*
e1	no	normal	yes
e2	yes	high	yes
e3	yes	very high	yes
e4	no	normal	no
e5	no	high	no
e6	yes	very high	no
e7	yes	high	no

Hence, we consider the value of *headache* random. Therefore, we can traditional statistic approach such as most frequent value replacement to decide that *headache* = *no*.

Example 2. In this example, the value of the *temperature* attribute of case *e*6 is missing shown in Table 6(a). Using the seven cases with complete data to rearrange the attribute so that temperature becomes the decision attribute, as shown in Table 6(b).

Selecting the attribute group (*headache*=*yes*, *flu*=*yes*, *temperature*=*normal*), the roughness measure is $\beta(T'_{temperature=normal}) = 0.67$. If the threshold is set at $b = 0.75$, the roughness measure $beta(T') < b$, considered small. We can then calculate the reduct set with these decision rules:

$$flu = yes \quad \rightarrow \quad temperature = high$$
$$flu = yes \quad \rightarrow \quad temperature = very\,high$$

Therefore, we can use either *high* or *very high* for the missing temperature value. Since *high* is the most frequent, the imputed value is determined to be *temperature* = *high*.

5. Related Work

For missing data imputation, there are enormous amount of work on ad hoc and statistic approaches in the literature, such as (Yuan 2010), but only a few methods were proposed using rough sets. So we shall briefly review some related work that used rough set for solving the missing data imputation problem.

Table 6. Imputation for missing value on *temperature*

(a) A value on *temperature* is missing

Case	Condition		Decision
	headache	*temperature*	*flu*
e1	yes	normal	no
e2	yes	high	yes
e3	yes	very high	yes
e4	no	normal	no
e5	no	high	no
e6	no	$*$	yes
e7	no	high	yes
e8	no	very high	no

(b) *temperature* as decision attribute

Case	Condition		Decision
	headache	*flu*	*temperature*
e1	yes	no	normal
e2	yes	yes	high
e3	yes	yes	very high
e4	no	no	normal
e5	no	no	high
e7	no	yes	high
e8	no	no	very high

Rough set approaches for handling missing values were introduced in 1990's (Grzymala-Busse & Wang 1997), (Kryszkiewicz 1998). Grzymala-Busse proposed rough set approaches to deal with three types of missing values: *loss values*, *attribute-concept values*, and "*do not care*" *conditions* (Grzymala-Busse & Grzymala-Busse 2007), (Grzymala-Busse, Grzymala-Busse, Hippe & Rząsa 2010).

The software toolkit Rough Set Exploration System (RSES) (Bazan & Szczuka 2005), developed by a team of researcher some of whom were involved in the original rough set theory research, uses the traditional approaches to deal with missing attribute values: removing objects with missing values, filling missing values with most common value (nominal) or the mean (numeric) of the attribute, treating missing value as information (null as regular value), and analysis using only the objects with complete data for reduct/rule calculation.

In (Latkowski 2005), the indiscernibility relation in rough set was enhanced to include individual treatment of missing values using two different approaches based on the assumption that not all missing values are semantically equal. An algorithm was provided in this study to create sub-optimal flexible indiscernibility relations for information with missing values.

A rough clustering approach dealing with missing data was proposed in (Li, Deogun, Spaulding & Shuart 2005). In this approach, traditional clustering techniques (such as K-means) was combined with soft computing (fuzzy and rough) to deal with the uncertainty in the data. It was reported in the study that rough K-means and fuzzy-rough K-means clustering algorithms yielded better performance.

An artificial neural network (ANN) approach was presented in (Setiawan, Venkatachalam & Hani 2008) that used rough set theory (RST) to reduce the dimensionality of the attributes through its reduct. Comparisons of the ANNRST (combination of ANN and RST) approach with other methods were given showing that the prediction accuracy using ANNRST was about the same as pure ANN without dimensionality reduction, and outperformed k-NN.

All of these methods kept the structure of the data (i.e. information table) with the original decision attribute unchanged. The method proposed in this paper differs from these approaches in a major way: the attribute with missing values is swapped with the original decision attribute so that the missing value can be "predicted" using the rules derived from rough set.

6. Conclusion

Rough set theory as a mathematical model for handling data with uncertainty has widely used in many application domains in the last two decades. The basic hypothesis of rough set theory is that the data set with uncertainty can be formally represented by a pair of approximations that are used to derive $condition \rightarrow$

decision rules. In this paper, we proposed the idea of rearrangement of attributes to explore the logical relations relations (may be considered "causal" relations) among the attributes. Roughness of rearrangements are calculated and optimal logical attribute flows are determined based on the roughness measures. With rearrangement of the missing-value-attribute becoming the decision attribute, the optimal logical attribute flows are used to determine if the decision rules deducted from rough set theory should be used for missing data imputation.

This paper is a preliminary study of the problem addressed. We are currently working on experiments of applying the method to real data sets, hopefully of relatively large sizes, and establishing evaluation criteria to measure the goodness of the imputation results.

References

Bazan, J. G. & Szczuka, M. (2005), The rough set exploration system, *in* 'Transactions on Rough Sets III', Springer, pp. 37–56.

Donders, A. R. T., van der Heijden, G. J., Stijnen, T. & Moons, K. G. (2006), 'Review: a gentle introduction to imputation of missing values', *Journal of clinical epidemiology* **59**(10), 1087–1091.

Grzymala-Busse, J. W. & Grzymala-Busse, W. J. (2007), An experimental comparison of three rough set approaches to missing attribute values, *in* 'Transactions on rough sets VI', Springer, pp. 31–50.

Grzymala-Busse, J. W., Grzymala-Busse, W. J., Hippe, Z. S. & Rząsa, W. (2010), 'An improved comparison of three rough set approaches to missing attribute values', *Control and Cybernetics* **39**(2), 469–486.

Grzymala-Busse, J. W. & Wang, A. Y. (1997), Modified algorithms lem1 and lem2 for rule induction from data with missing attribute values, *in* 'Proc. of the Fifth International Workshop on Rough Sets and Soft Computing at the Third Joint Conference on Information Sciences', Research Triangle Park, NC, pp. 69–72.

Kryszkiewicz, M. (1998), 'Rough set approach to incomplete information systems', *Information sciences* **112**(1), 39–49.

Latkowski, R. (2005), 'Flexible indiscernibility relations for missing attribute values', *Fundamenta informaticae* **67**(1), 131–147.

Li, D., Deogun, J., Spaulding, W. & Shuart, B. (2005), Dealing with missing data: Algorithms based on fuzzy set and rough set theories, *in* 'Transactions on Rough Sets IV', Vol. 3700 of *Lecture Notes in Computer Science*, Springer, pp. 35–57.

Pawlak, Z. (1982), 'Rough set', *International Journal of Parallel Programming* **11**(5), 341–356.

Pawlak, Z., Grzymala-Busse, J., Slowinski, R. & Ziarko, W. (1995), 'Rough sets', *Communication of the ACM* **38**(11), 89–95.

Pawlak, Z. & Skowron, A. (2007), 'Rudiments of rough sets', *Information Sciences* **177**(1), 3–27.

Rubin, D. B. (2009), *Multiple imputation for nonresponse in surveys*, Vol. 307, Wiley.com.

Schafer, J. L. (2010), *Analysis of incomplete multivariate data*, CRC press.

Setiawan, N. A., Venkatachalam, P. & Hani, A. F. M. (2008), Missing attribute value prediction based on artificial neural network and rough set theory, *in* 'Proceedings of International Conference on BioMedical Engineering and Informatics', Vol. 1, IEEE, pp. 306–310.

Yuan, Y. C. (2010), 'Multiple imputation for missing data: Concepts and new development (version 9.0)', *SAS Institute Inc, Rockville, MD* .