

Performance Comparison between Two Interpretations of Missing Data using Matrix-Characterized Approximations

Author(s): ¹Thin Thin Soe, ²Myat Myat Min

Affiliation(s): ¹Web Mining Lab, ²Faculty of Computer Science

University of Computer Studies, Mandalay, Myanmar

*Corresponding Author: thinthinsoe@ucsm.edu.mm

ORIGINAL
ARTICLE



Abstract - Nowadays, the veracity related with data quality such as incomplete, inconsistent, vague or noisy data creates a major challenge to data mining and data analysis. Rough set theory presents a special tool for handling the incomplete and imprecise data in information systems. In this paper, rough set based matrix-represented approximations are presented to compute lower and upper approximations. The induced approximations are conducted as inputs for data analysis method, LERS (Learning from Examples based on Rough Set) used with LEM2 (Learning from Examples Module, Version2) rule induction algorithm. Analyses are performed on missing datasets with “do not care” conditions and missing datasets with lost values. In addition, experiments on missing datasets with different missing percent by using different thresholds are also provided. The experimental results show that the system outperforms when missing data are characterized as “do not care” conditions than represented as lost values.

Keywords: *rough set, incomplete data, missing values, matrix-represented approximations, “do not care” conditions, lost values*

I. INTRODUCTION

Available knowledge about the real world is inherently uncertain, and decisions have been usually made based on incomplete and partially imprecise data. The incomplete data means that data in which some features are missing from its particular features. The occurrence of incomplete data in either the testing set or training set affects the learning quality of the classifiers. Since rough set theory

(RST) is a special tool for handling the imprecise and incomplete data in information systems, many researchers have presented rough set based methods for handling incomplete data [1, 5, 7, 9, 10, 12]. A sequential matrix-based algorithm (SMA) which calculates lower and upper approximations of incomplete information systems is introduced in [2]. However, they did not mention about handling missing datasets with different missing percent.

The speeding up incomplete data analysis system using matrix-represented approximations is proposed in [11]. This system enables to handle missing data within the acceptable time and speedup than the traditional method. In which, missing datasets with different missing percent are examined by different thresholds. Also, the performance comparison between the traditional rough set and the system is conducted. The two types of missing attribute values: “do not care” conditions and lost values are also depicted. However, the performance comparison between the two types of missing values was not mentioned specifically. Therefore, in this paper, we contribute that;

- A bunch of experiments on five missing datasets wherein missing values are represented as “do not care” conditions by using different thresholds
- A set of analyses on five missing datasets where missing values are represented as lost values by using different thresholds
- The performance comparison between two characterizations of missing values: “do not care” conditions and lost values

The rest of the paper is arranged as follows. Section 2 expresses the existing methods. The basic concept of incomplete data analysis with matrix-represented approximations and the case study of the system are presented in Section 3. Experimental results with two types

of missing values and performance comparison are discussed in Section 4. Finally, the paper ends with conclusions in Section 5.

II. RELATED WORK

Rough set based characteristic relation for incomplete decision tables, was introduced by Grzymala [5]. According to [1, 7, 9, 10], there were three main characterizations of missing attribute values: “do not care” conditions, lost values and attribute-concept values. Rough set approach to missing attribute values with “do not care” conditions was proposed in [9, 10]. In this approach, each missing attribute value was replaced by all possible values of that attribute. Missing attribute values represented as lost values (i.e. the original value was erased) was presented in [1]. Another approach with attribute-concept values was described in [7].

The matrix characterizations of the lower and upper approximations in set-valued information systems and two incremental approaches for updating the relation matrix were introduced in [3]. Authors proposed a sequential matrix-based algorithm (SMA) and three parallel methods based on MapReduce to calculate approximations in incomplete decision tables [2]. SMA is a sequential matrix-based algorithm which computes the approximations of the decision, the positive region, the negative region and the boundary region.

In [11], the speeding up incomplete data analysis system using matrix-represented approximations was proposed. Moreover, by different thresholds, a set of experiments on datasets with different missing percent was implemented. Also, the performance comparison between the traditional rough set and the system was presented.

III. MISSING DATA ANALYSIS WITH MATRIX-REPRESENTED APPROXIMATIONS

The characteristic sets for the incomplete decision table are calculated initially. Based on the resulting characteristic sets, matrix-represented lower and upper approximations are generated.

The detailed descriptions of these steps are presented in Section 3.2 and Section 3.3 respectively. The induced lower and upper approximations are used as inputs for data analysis method, LERS (Learning from Examples based on Rough Set) used with LEM2 (Learning from Examples Module, Version2) rule induction algorithm [4, 8].

A. Missing Data

In *RST*, a decision table is exploited to describe an information system [5, 6, 7]. Each row of the decision table corresponds to a case and columns stand for attributes (a finite set of condition attributes and a decision attribute). The set of all cases and the set of all attributes are presented by U and A respectively. Then the value of an attribute ‘ a ’ for a case ‘ c ’ is specified as $a(c)$.

A decision table is incomplete when there are some missing attribute values. In this paper, two main types of missing values: “do not care” conditions ‘ $*$ ’ and lost values ‘?’ are presented. Table I shows a sample missing dataset with lost values ‘?’; in which all missing values can be represented as lost values ‘?’, or can be represented as “do not care” conditions ‘ $*$ ’. The complete information of the attribute values is depicted in [13].

B. Characteristic Relation

In *RST*, a decision table is exploited to describe an information system [5, 6, 7]. Each row of the decision table corresponds to a case and columns stand for attributes (a finite set of condition attributes and a decision attribute). The set of all cases and the set of all attributes are presented by U and A respectively. Then the value of an attribute ‘ a ’ for a case ‘ c ’ is specified as $a(c)$.

The characteristic relation, a generalization of indiscernibility relation, is used to describe incompletely specified tables. The characteristic set $K_A(c)$ is the set of all cases U that are indistinguishable from ‘ c ’ using all attributes A [2, 5].

$$K_A = \{ (c_1, c_2) \mid \forall a \in A, (a(c_1) \neq '?') \wedge (a(c_1) = a(c_2) \vee a(c_1) = '*' \vee a(c_2) = '*') \} \quad (1)$$

Based on the characteristic set, the characteristic relation $R(A)$ is defined as follows.

$$(c_1, c_2) \in R(A) \Leftrightarrow c_2 \in K_A(c_1) \quad (2)$$

TABLE I. SAMPLE MISSING DATASET WITH LOST VALUES ‘?’

Case	Cap-shape	Cap-surface	Cap-color	Bruises	Odor	Gill-attachment	Gill-spacing	Gill-size	Gill-color	Stalk-shape	Stalk-root	Stalk-surface-above-ring	Stalk-surface-below-ring	Stalk-color-above-ring	Stalk-color-below-ring	Veil-type	Veil-color	Ring-number	Ring-type	Spore-print-color	Population	Habitat	Decision
1	x	y	y	t	l	f	c	b	g	e	c	s	s	w	w	p	w	o	p	n	n	g	e
2	x	?	y	t	?	f	c	b	?	e	?	s	s	w	w	p	w	o	p	k	s	?	e
3	b	s	?	t	a	f	c	b	w	e	c	s	s	?	w	p	w	o	p	n	s	g	e
4	?	f	n	f	?	?	w	b	n	t	e	?	f	w	w	p	w	o	e	k	a	g	e
5	b	y	w	t	a	f	c	b	w	e	c	s	s	w	w	p	w	o	p	n	n	m	e
6	x	y	?	t	p	f	c	n	?	e	e	s	s	w	w	p	w	o	p	n	v	u	p
7	x	s	?	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	k	s	g	p
8	x	y	w	t	p	f	c	n	n	e	?	s	s	w	w	?	w	o	p	n	s	u	p
9	?	?	?	t	?	f	c	?	?	e	?	s	s	w	w	p	w	o	p	n	?	?	p
10	x	?	n	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	n	v	g	p

Firstly, the characteristic sets for Table I with “do not care” conditions “*” are calculated by using the equation (1).

$$K_A(10) = \{10\}$$

$$K_A(1) = \{1, 9\}$$

$$K_A(2) = \{2\}$$

$$K_A(3) = \{3, 9\}$$

$$K_A(4) = \{4\}$$

$$K_A(5) = \{5, 9\}$$

$$K_A(6) = \{6, 9\}$$

$$K_A(7) = \{7\}$$

$$K_A(8) = \{8, 9\}$$

$$K_A(9) = \{1, 3, 5, 6, 8, 9, 10\}$$

$$K_A(10) = \{9, 10\}$$

Then, the characteristic sets for Table I with lost values ‘?’ are computed as follows.

$$K_A(1) = \{1\}$$

$$K_A(2) = \{2\}$$

$$K_A(3) = \{3\}$$

$$K_A(4) = \{4\}$$

$$K_A(5) = \{5\}$$

$$K_A(6) = \{6\}$$

$$K_A(7) = \{7\}$$

$$K_A(8) = \{8\}$$

$$K_A(9) = \{1, 5, 6, 9, 10\}$$

C. Matrix-Represented Approximations

The calculation of lower and upper approximations is an essential part in rough set-based knowledge acquisition systems. Among the definitions of approximations [6], concept lower and upper approximations are utilized in this paper. A concept X means that the set of all cases or examples with the same decision value. The lower approximation is the set of all cases which are classified as members of the concept X . Then the upper approximation contains the set of cases which can be possible members of the concept X . The two concepts of Table I are $X_1 = \{1, 2, 3, 4, 5\}$ and $X_2 = \{6, 7, 8, 9, 10\}$.

Firstly, the relation matrix RM of the incomplete decision table is generated based on the characteristic relation [11]. The relation matrix RM , an $n \times n$ matrix representing K_A , is

$$RM_{n \times n}^{K_A} = (m_{ij})_{n \times n} \quad (3)$$

Where,

n = number of cases,

$$1 \leq i, j \leq n,$$

$$m_{ij} = \begin{cases} 1, (c_i, c_j) \in K_A \\ 0, (c_i, c_j) \notin K_A \end{cases}$$

$$m_{ii} = 1$$

And then, the induced diagonal matrix IDM is constructed through the relation matrix. The induced diagonal matrix IDM is denoted as follows.

$$IDM_{n \times n}^{KA} = \text{diag} \left(\frac{1}{\sum_{j=1}^n m_{1j}}, \frac{1}{\sum_{j=1}^n m_{2j}}, \dots, \frac{1}{\sum_{j=1}^n m_{nj}} \right) \quad (4)$$

Where,

n = number of cases, $1 \leq j \leq n$

The decision matrix DM is computed according to the concept X . The decision matrix DM is expressed as:

$$DM_{n \times r}^X = (G(X_1), G(X_2), \dots, G(X_r)) \quad (5)$$

Where,

r = number of concepts

n = number of cases

$$G(X_1) = (g_1, g_2, \dots, g_n)^T$$

$$g_i = \begin{cases} 1, & (c_i) \in X \\ 0, & (c_i) \notin X \end{cases}$$

Through the matrix multiplication of the induced diagonal matrix, the relation matrix and the decision matrix, the basic matrix BM is calculated. The basic matrix BM :

$$BM(X) = IDM_{n \times n}^{KA} \cdot RM_{n \times n}^{KA} \cdot DM_{n \times r}^X \quad (6)$$

The resulting basic matrix $BM(X)$ let be $(b_1, b_2, \dots, b_n)^T$. Then, the matrix-represented lower and upper approximations are computed as follows.

$$\underline{AX} = BM^{[\alpha, 1]}(X) \quad (7)$$

$$\overline{AX} = BM^{(\beta, 1]}(X) \quad (8)$$

Where,

$$BM^{[\alpha, 1]}(X) = (b_i')_{n \times 1}$$

$$b_i' = \begin{cases} 1, & \alpha \leq b_i \leq 1 \\ 0, & \text{else} \end{cases}$$

$$BM^{(\beta, 1]}(X) = (b_i')_{n \times 1}$$

$$b_i' = \begin{cases} 1, & \beta < b_i \leq 1 \\ 0, & \text{else} \end{cases}$$

n = number of cases, $1 \leq i \leq n$,

$$0 \leq \alpha \leq 1, 0 \leq \beta \leq 1$$

In this paper, different values of α and β are exploited instead of using only $\alpha = 1$ and $\beta = 0$ [2]. This threshold did not affect the performance of the system examined on missing datasets with smaller amount of missing attributes values. The accuracy is lower than the traditional rough set approach whereas analyzing missing datasets with more missing values. The examining results and discussions were

described in our previous work [11]. In this paper, we emphasize that the performance comparison between the two representations of missing values, “do not care” conditions and lost values. The examined results have been depicted in the Section 4.

For the sample missing dataset, illustrated in Table I, matrix-representing lower and upper approximations for both types of missing values are calculated as follows. Firstly, the relation matrix RM is constructed with regard to the characteristic relation. Based on the resulting relation matrix RM , the induced diagonal matrix IDM is computed. Then, the decision matrix DM is calculated via the concept X . The basic matrix BM is constructed through the matrix multiplication of IDM , RM and DM . The lower and upper approximations for table 1 interpreted as “do not care” conditions are computed using the basic matrix BM with $\alpha = 1$ and $\beta = 0$. The evaluated result is depicted in Table II. In Table III, the basic matrix, the matrix-represented lower and upper approximations for sample missing dataset with lost values are illustrated.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this experiment, mushroom dataset from UCI Machine Learning Repository [13] is exploited. Afterwards, five missing datasets are generated by assigning different amounts (10%, 15%, 20%, 25% and 30%) of missing attribute values to this dataset. This experiment is coded in JAVA and performed on an Intel core i5 processor for Windows 7, 2GB RAM and 500GB hard disk. The performance of the system is assessed with the accuracy and the execution time.

The resultant five missing datasets are initially interpreted as missing datasets with “do not care” conditions ‘*’. Firstly the 10% missing dataset is examined with different thresholds, $(\alpha = 0.8, \beta = 0.1)$, $(\alpha = 0.8, \beta = 0.2)$, $(\alpha = 0.9, \beta = 0.1)$, $(\alpha = 0.9, \beta = 0.2)$ and $(\alpha = 1, \beta = 0)$. In this experiment, the accuracy remains the same for all thresholds. And then, the 15%, 20%, 25% and 30% missing datasets are analyzed with different thresholds. For the 15% and 20% missing datasets, the accuracy is the same as analyzing with the 10% missing dataset. In experiment with the 25% missing dataset, the accuracy decreases when $(\alpha = 1, \beta = 0)$. For the 30% missing dataset, the accuracy decreases when $(\alpha = 0.9, \beta = 0.1)$, $(\alpha = 0.9, \beta = 0.2)$ and $(\alpha = 1, \beta = 0)$. The experimental result is depicted in Fig. 1.

Then, the five missing datasets are interpreted as missing datasets with lost values ‘?’. Each of these missing datasets is analyzed with different thresholds. Analyzing the 10%, 15%, 20% and 25% missing datasets shows that the accuracy remains the same for all thresholds. In examining with the 30% missing dataset, the accuracy decreases for all different thresholds. The comparison of the accuracy on these missing datasets with different thresholds is illustrated in Fig. 2.

TABLE II. MATRIX-REPRESENTED APPROXIMATIONS FOR TABLE I WITH “DO NOT CARE” CONDITIONS

Basic Matrix BM		Lower Approximation $BM [1,1]$		Upper Approximation $BM (0,1]$	
$X1$	$X2$	$X1$	$X2$	$X1$	$X2$
0.50	0.50	0	0	1	1
1.00	0.00	1	0	1	0
0.50	0.50	0	0	1	1
1.00	0.00	1	0	1	0
0.50	0.50	0	0	1	1
0.00	1.00	0	1	0	1
0.00	1.00	0	1	0	1
0.00	1.00	0	1	0	1
0.43	0.57	0	0	1	1
0.00	1.00	0	1	0	1

TABLE I. MATRIX-REPRESENTED APPROXIMATIONS FOR TABLE I WITH LOST VALUES

Basic Matrix BM		Lower Approximation $BM [1,1]$		Upper Approximation $BM (0,1]$	
$X1$	$X2$	$X1$	$X2$	$X1$	$X2$
1.00	0.00	1	0	1	0
1.00	0.00	1	0	1	0
1.00	0.00	1	0	1	0
1.00	0.00	1	0	1	0
1.00	0.00	1	0	1	0
0.00	1.00	0	1	0	1
0.00	1.00	0	1	0	1
0.00	1.00	0	1	0	1
0.40	0.60	0	0	1	1
0.00	1.00	0	1	0	1

After examining with both types of missing values, we found that the accuracy remains the same for all thresholds while using missing datasets with smaller missing percent. For the 25% missing dataset with the threshold ($\alpha = 1$, $\beta = 0$), the system outperforms when missing values are characterized as lost values than represented as “do not care” conditions. However, for the 30% missing dataset with the threshold ($\alpha = 1$, $\beta = 0$), the accuracy remains the same for both interpretations of missing values. For both datasets with larger missing percent, the system outperforms when missing values are represented as “do not care” conditions than represented as lost values while using ($\alpha = 0.8$, $\beta = 0.1$) and ($\alpha = 0.8$, $\beta = 0.2$).

The computational time on missing datasets with “do not care” conditions and the execution time on missing datasets with lost values are compared in Fig. 3. In this experiment, different data sizes (100, 2000, 5000) are used and the threshold ($\alpha = 1$, $\beta = 0$) is used for both missing values. As shown in Fig. 3, the execution time on both missing values are not significantly different up to 2000 records. For the missing dataset with 5000 records, the lost value interpretation is slightly faster about 29.379 seconds than the representation of “do not care” conditions.

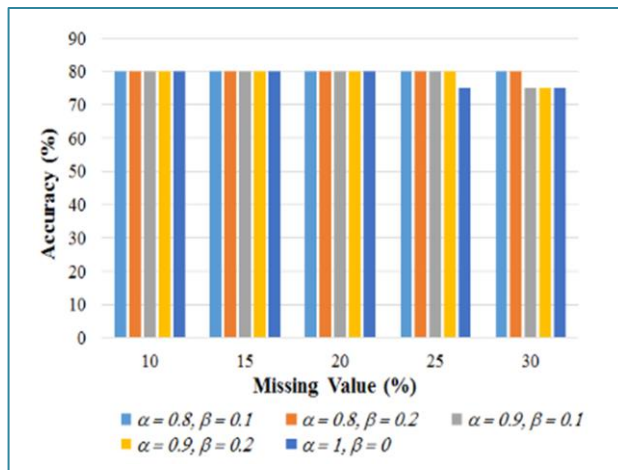


Fig. 1. Experimental results for missing datasets with “do not care” conditions using different thresholds.

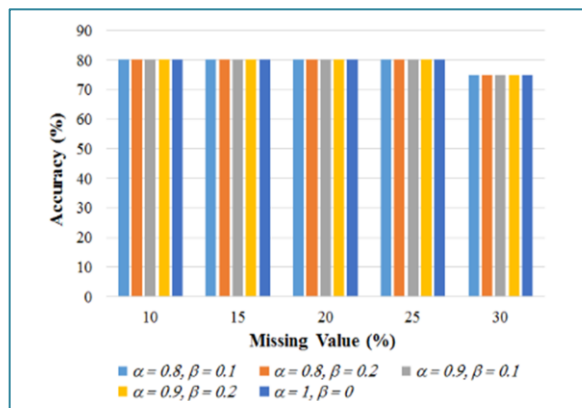


Fig. 2. Experimental results for missing datasets with lost values using different thresholds.

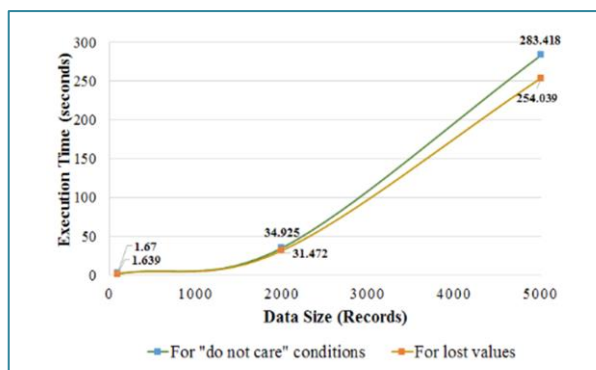


Fig. 3. Comparison of execution time between the two interpretations of missing data on different data sizes.

V. CONCLUSIONS

At present, data in many real-life applications are incomplete, inconsistent, vague or noisy due to inherent measurement inaccuracies or intentional blurring of data. Incomplete data handling is an important issue due to the incomplete data in either the testing set or training set affects the learning quality of the classifiers. In this paper,

by using different thresholds, we first presented a set of experiments on missing datasets with “do not care” conditions ‘*’. Then, evaluating on missing datasets with lost values ‘?’ was provided. According to the experimental results, the accuracy on the datasets with smaller missing percent remains the same for all thresholds for both types of missing values. With the thresholds ($\alpha=0.8, \beta=0.1$) and ($\alpha=0.8, \beta=0.2$), the system outperforms when missing values are represented as “do not care” conditions ‘*’ than it is represented as lost values ‘?’. The lost value interpretation is slightly faster than the representation of “do not care” conditions for larger datasets. The execution time on both missing values are not significantly different for smaller datasets. Then evaluating on larger data sets will be performed in advance using MapReduce based matrix-represented approximations.

DECLARATION

The authors have disclosed no conflicts of interest and the project was self-funded.

REFERENCES

- [1] J. Stefanowski, and A. Tsoukiàs, “On the Extension of Rough Sets Under Incomplete Information”, *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*, Springer, Berlin, Heidelberg, 1999, pp. 73-81.
- [2] J. Zhang, J. S. Wong, Y. Pan, and T. Li, “A Parallel Matrix-Based Method for Computing Approximations in Incomplete Information Systems”, *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, 2015, pp. 326-339.
- [3] J. Zhang, Li. Tianrui, D. Ruan, and D. Liu, “Rough Sets based Matrix Approaches with Dynamic Attribute Variation in Set-Valued Information Systems”, *International Journal of Approximate Reasoning*, vol. 53, 2012, pp. 620-635.
- [4] J.W. Grzymala-Busse, “LERS-A System for Learning from Examples based on Rough Sets”, *Intelligent Decision Support*, vol. 11, Springer, Dordrecht, 1992, pp. 3-18.
- [5] J.W. Grzymala-Busse, “Characteristic Relations for Incomplete Data: A Generalization of the Indiscernibility Relation”. *Transactions on rough sets IV*, Springer, Berlin, Heidelberg, 2005, pp. 58-68.
- [6] J.W. Grzymala-Busse, “Rough Set Strategies to Data with Missing Attribute Values”, *Foundations and Novel Approaches in Data Mining*. Studies in Computational Intelligence, vol. 9, Springer, Berlin, Heidelberg, 2005, pp. 197-212.
- [7] J.W. Grzymala-Busse, “Three Approaches to Missing Attribute Values: A Rough Set Perspective”, *Data Mining: Foundations and Practice*. Studies in

Computational Intelligence, vol. 118, Springer, Berlin, Heidelberg, 2008, pp. 139-152.

- [8] J.W. Grzymala-Busse, and B. W. Chien Pei, "Classification Methods in Rule Induction", *Proc of the Fifth Intelligent Information Systems Workshop*, Deblin, Poland, 1996.
- [9] M. Kryszkiewicz, "Rough Set Approach to Incomplete Information Systems", *Information sciences*, 112(1-4), 1998, pp. 39-49.
- [10] M. Kryszkiewicz, "Rules in Incomplete Information Systems", *Information sciences*, 113(3-4), 1999, pp. 271-292.
- [11] T.T.Soe, and M.M.Min, "Speeding up Incomplete Data Analysis using Matrix-Represented Approximations", 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (IEEE/ACIS SNPD 2018), Busan, Korea, June 27-29, 2018, pp. 206-211.
- [12] Z. Pawlak, "Rough Sets: Theoretical Aspects of Reasoning about Data", *System Theory*, Kluwer Academic Publishers, Boston, London, Dordrecht, 1991. <https://archive.ics.uci.edu/ml/datasets/Mushroom>