Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 11, Number 3 (2015), pp. 1147-1155

http://www.ripublication.com

Handling Missing Value on Meteorological Data Classification with Rough Set Based Algorithm

Winda Aprianti[1*] and Imam Mukhlash[2]

[1,2]Department of Mathematics, Institut Teknologi Sepuluh Nopember Surabaya, Indonesia
[1]winda.ap17@gmail.com [2]imamm@matematika.its.ac.id

Abstract

Data mining is a process to find patterns and knowledge from a large database. One task in data mining is classification, which is the process of finding rules for predicting the object in the database. The weather has an important role in human life, such as in the areas of social and economic welfare, agriculture, disaster management, and finance. So, weather prediction needs to make planning in various fields. Most of the database cannot be separated from the incompleteness problem, which is caused by faulty procedures manual data entry, incorrect measurements, equipment faults, and many others. In this research, we use incomplete meteorological dataset. Before applying the rough set to obtain rules, incomplete dataset is converted into a complete dataset by replacing the missing value with the average value of the records that have the same decision class. Then we find the lower and upper approximation. We obtained certain rules from lower approximation and possible rules from upper approximation. To test the performance of this algorithm, we applied rules to test data. Result of the application of this algorithm on datasets contain 5%, 10%, 15%, 20%, 25%, and 30% missing value show that increasing of missing value lead to the accuracy of rules decreases and the number of rules no affects the accuracy of rules. Resulted rules by rough set algorithm effectively to predict rainfall for dataset contain missing value less than 25%.

Keyword: meteorological, incomplete dataset, rough set, classification

INTRODUCTION

The development of database technology is rapidly that causes the storage of data from various sources can be done easily and quickly. As a result, the volume of data

generated every day increased so analyze that data becomes an important requirement. Data mining is a method to discover knowledge in a large database in the data. One task in data mining is a classification, which is the process of finding rules for predicting the object in the database [1]. Datasets are often used for classification are meteorological data. This is due to the weather that has an important role in human life, such as in the areas of social and economic welfare, agriculture, disaster management, and finance [2]. So, weather prediction needs to make planning in various fields. Meteorological dataset is uncertain, in which the two objects have the same value for all attributes of a class may have two different decisions.

Most of the database cannot be separated from the incompleteness problem, which is caused by faulty procedures manual data entry, incorrect measurements, equipment faults, and many others [3]. This also occurs in meteorological data. Aprianti and Mukhlash [4] applied rough set and fuzzy rough set algorithm on incomplete meteorological dataset directly, without change it into a complete dataset. As a result, rules generated by rough set algorithm cannot predict some conditions, while rules based on fuzzy rough set algorithm can predict all condition. In contrast to [4], in this paper we will handle missing value by replacing the missing value with the average value of the attributes of the object that has the same decision class. Other than that, we will look for plausibility value of each rules generated by rough set algorithm.

Incomplete datasets can be converted into a complete data in various ways, Chmielewski et al [5] removed object with unknown values before beginning the process of learning, whereas Mukhlash et al [6] records that contain missing values replaced with the average value of the attribute other. However, deleting the record causes some information is lost. In Wang and Fan [7], a missing value is replaced with the most frequently value or average value of the records that have the same decision class.

Therefore, we apply rough set algorithm in incomplete meteorological datasets that has been preprocessing to obtain certain and possible rules. Rules that have been obtained will be used to predict weather condition.

MATERIALS AND METHODS

Data

In this research, we use secondary weather data in [6]. Attributes of the data used is average of temperature, relative humidity, air pressure and speed of wind as a supporting attribute, and rainfall as the decision attribute. Incomplete datasets obtained by eliminating the supporting attribute values of multiple objects at random, hereinafter referred to as incomplete decision table.

Rough Set

Rough set theory (RST) is an extension of conventional set theory that supports approximations in decision making. A rough set is itself the approximation of a vague concept (set) by a pair of precise concepts, called lower and upper approximations [8].

Some basic concepts in rough set theory are information systems, equivalence relation, lower approximation, and upper approximation. $I = (U, A)$ is an information system, where U is a non-empty set of finite objects and A is a non-empty finite set of attributes so that $a: U \rightarrow V_a$ for every $a \in A$. V_a is the set of values that attribute a may take. For decision system, $A = \{C \cup D\}$ where C is the set of input features and D is the set of class indices [8].

With any $A_j \subseteq A$ there is an associated equivalence relation:

$$IND(A_j) = \{(Obj^{(i)}, Obj^{(k)}) \in U \times U \mid a \in A_j, a(Obj^{(i)}) = a(Obj^{(k)})\}$$

Note that this corresponds to the equivalence relation for which two objects are equivalent if they have the same value of attribute A_j, $Obj^{(i)}$ and $Obj^{(k)}$ are said to have an indiscernibility relation on attribute A_j. The equivalence classes of the indiscernibility relation with respect to A_j are denoted by $[Obj^{(i)}]\, A_j$ [8].

Let X is subset of universe U and A_j is a subset arbitrary of attribute A. The lower and upper approximation A_j on X, respectively defined as follows.

$$A_j X = \{Obj^{(i)} \mid Obj^{(i)} \in U, [Obj^{(i)}]A_j \subseteq X\}, \text{ and}$$

$$\overline{A_j} X = \{Obj^{(i)} \mid Obj^{(i)} \in U \text{ and } [Obj^{(i)}]A_j \cap X \neq \emptyset\}$$

Methodology

The framework of this research conducted in several steps, i.e. data preprocessing, application of rough set algorithm, and algorithm testing.

The first step is preprocessing of the incomplete dataset. This step is done by replacing the missing value in the attribute with the average value of the records that have same decision class. So, we get complete decision table.

Rules are obtained by applying rough set algorithm to complete decision table. Firstly, we find equivalence class of each attribute. Secondly, we find the lower and upper approximation from equivalence class. For the upper approximation, we calculate plausibility of each upper approximation. Plausibility is ratio of the number of object corresponding to attribute values and decision class with the number of object corresponding to attribute values. We will get on certain rules from the lower approximation, while the possible rules obtained from the upper approximation. Finally, to test the performance of the rules generated by rough set algorithm, we apply rules to predict test data.

RESULT AND DISCUSSION

Classification Algorithm

The basic idea of classification algorithms in this paper refers to the study [9] on the approximation of rough sets in incomplete information systems, studies [10] on the use of fuzzy set theory and the variable precision rough set to discover fuzzy knowledge from quantitative data, and [11] on the use of fuzzy rough sets of quantitative data is incomplete. Based on these studies, we propose algorithm by

modifying algorithm in [9] and [10] by changing the missing value prediction process in step 2 and the calculation of plausibility in step 5. This algorithm can be formed as follows:

Step 1: Partition the set of objects into disjoint subsets according to the label. Denote each set of objects belonging to the same class C_l as X_l.
Step 2: Transformation quantitative values into categorical. If the *object (i)* has a missing value for the attribute A_j, missing value will be replaced with the average value of the records that have same decision class.
Step 3: Find equivalence classes of a single attribute.
Step 4: Initialize $q = 1$, where q is used to calculate the number of attributes when processed for the lower and upper approximations.
Step 5: Calculate the lower and upper approximations of each subset B with q attributes for each class X_l. Then, it will look for the value of plausibility of each upper approximation.
Step 6: Set $q = q + 1$ and repeat steps 5-7, until $q > m$
Step 7: Generate certain rules of the lower approximation and upper approximation of possible rules in each subset B.
Step 8: Remove certain and possible rules with conditions more specific sections and has a value less than or equal plausibility than others on certain and possible rules.
Step 9: Output on certain and possible rules.

Testing and Result Analysis

As an example of the classification process, we use the secondary data obtained from [6], taken 50 dataset as a dataset in this study. The dataset is shown in Table 1.

Table 1 Incomplete Meteorological Dataset

Object	Average Temperature (A)	Relative Humidity (B)	Air Pressure (C)	Speed of Wind (D)	Rainfall (RF)
1.	25.4	96.4	1007	7.6	45
2.	26.8	93.5	1007.9	12.2	8.9
3.	27	*	1008.6	12.6	8.9
4.	27.7	91.2	*	12.2	7.1
5.	27.8	89	1008.7	13.8	0
⋮	⋮	⋮	⋮	⋮	⋮
46.	*	96.3	1010	9.7	0
47.	29.2	88.2	1008.1	6.6	0
48.	29.1	*	1008.5	7.2	38.1
49.	28.4	87.6	1007.3	*	0
50.	25.2	97.1	1007.4	3.7	37.1

Before applying the rough set algorithm, we perform preprocessing to the dataset in Table 1 by replacing the missing value with the average value of the records that have same decision class. After the complete dataset, all of the attributes converted into categorical data.

For average of temperature (*A*), when *A* less than 26.5 is changed to cold, the *A* between 26.5 and 29 changed to normal, and *A* more than 29 changed into hot. For the relative humidity (*B*), when *B* less than 68 was changed to dry, *B* between 68 and 78 changed to humid, and *B* more than 78 changed to wet. For air pressure (*C*), when *C* less than in 1008 changed into low, *C* between 1008 and 1013 changed into medium, and *C* more than 1013 changed into high. For speed of wind (*D*), when *D* less than 4 transformed into slow, *D* between 4 and 8 changed into normal, and *D* more than 8 changed into windy. Based on criteria rainfall intensity per day of World Meteorological Organization, the decision attribute or rainfall (*RF*) is given categories as follows: if *RF* less than 5 changed to Very Smooth Rain (VSR), if RF is between 5 and 20 changed to Smooth Rain (SR), if RF is between 20 and 50 changed to Normal (N), if RF is between 50 and 100 changed to Heavy Rain (HR), and if more than 100 RF changed to Heavier Rain (HrR).

The object 1 to object 40 is used as training data, while the object 41 to object 50 is used as the test data. Rough set algorithm is applied to the training data set to produce 54 rules as shown in Table 2.

Table 2 Rules Generated by Rough Set Algorithm

No.	Rules
	Certain Rules
1.	If C = Medium, D = Slow, then RF = N
2.	If A = Cold, C = Medium, D = Normal, then RF = N
3.	If A = Normal, C = Medium, D = Slow, then RF = N
Possible Rules	
4.	If A = Cold, then RF = VSR, p = 0.29
⋮	⋮
53.	If A = Cold, C = Low, D = Windy, then RF = N, p = 0.5
54.	If B = Wet, C = Low, D = Slow, then RF = N, p = 0.25

Note: p = plausibility

We use the results in Table 2 rules to predict rainfall in the test data. If there are a few rules that satisfy the conditions of testing the data, then we took the decision with the largest average value. The predicted results are shown in Table 3.

Table 3 Rainfall Prediction of Test Data

Object	Rainfall
41.	SR
42.	SR
43.	SR
44.	SR
45.	N
46.	SR
47.	SR
48.	SR
49.	SR
50.	N

The results of the comparison between the object 41 until 50 in Table 1 and Table 3, shows that there are 3 results different predictions, i.e. the object 43, 46, and 50. Thus, obtained the accuracy is 70%.

To test the algorithm, we apply rough sets classification algorithm to the incomplete dataset consisting of 1735 objects, each of which contains a missing value by 5%, 10%, 15%, 20%, 25%, and 30%. Selection of training and testing data performed by 10-fold validation.

The steps of generate rules on each dataset same as generate rules to training data in Table 1. Application of rough set algorithm on each dataset is done by implementing rough sets algorithm in Matlab. From the application of rough set algorithm on each dataset obtain the number of rules which are presented in Table 4.

Table 4 Number of Rules Generated by Rough Set Algorithm

k	5%	10%	15%	20%	25%	30%
1	26	27	27	27	27	27
2	26	27	26	27	27	34
3	26	27	27	27	27	27
4	26	34	27	27	27	27
5	26	27	27	27	26	27
6	26	27	27	27	27	27
7	26	27	27	27	27	27
8	26	27	27	26	27	26
9	25	26	27	27	27	27
10	26	26	27	27	27	27
Average	26	28	27	27	27	28

Rules generated by rough set algorithm are used to predict rainfall in the test data. Then result of predicted rainfall compared with the actual rainfall to obtain the accuracy of rules. The results of the accuracy of rules generated by rough set algorithm are presented in Table 5 and represented by Figure 1.

Table 5 Accuracy of Rules Generated by Rough Set Algorithm

k	5%	10%	15%	20%	25%	30%
1	85.55%	78.74%	73.56%	73.99%	71.84%	66.47%
2	85.55%	85.06%	78.16%	69.94%	67.24%	63.79%
3	79.31%	78.03%	76.44%	73.41%	73.41%	70.69%
4	84.97%	78.03%	77.01%	70.11%	69.94%	65.52%
5	83.33%	74.14%	73.41%	78.74%	72.25%	69.94%
6	87.36%	83.82%	73.99%	70.55%	71.10%	67.63%
7	85.06%	81.04%	71.84%	71.68%	70.11%	75.72%
8	81.50%	83.24%	73.99%	73.56%	67.24%	71.68%
9	86.13%	79.89%	74.57%	72.99%	66.67%	71.26%
10	85.06%	75.72%	82.66%	72.99%	67.05%	72.99%
Average	84.38%	79.769%	75.56%	72.80%	69.69%	69.57%

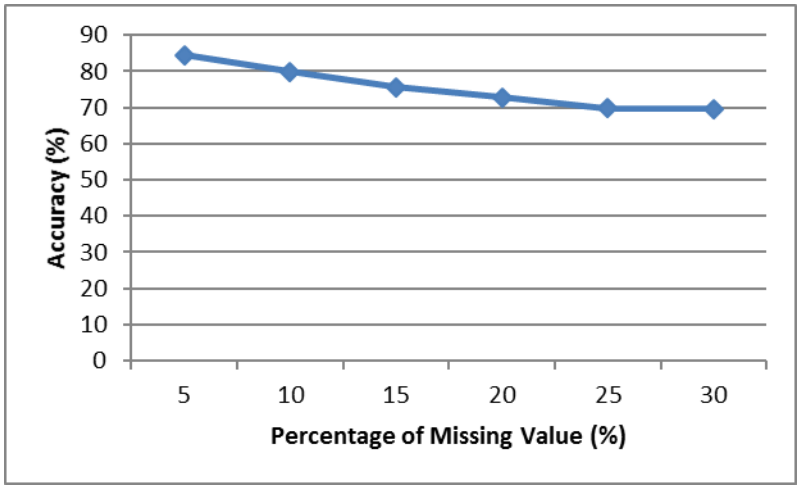

Figure 1 Graph of Comparison Rules Accuracy with Percentage Missing Value

Figure 1 show that increasing missing value causes the accuracy of rules decreases. When percentage of missing value between 5% and 20% accuracy is above 70%, but when the missing value increased to 25% and 30% accuracy is below 70%. This indicates that the rules generated by on rough sets algorithm have a high accuracy for the missing value below 25%. Figure 2 show that the number of rules no affects the accuracy of rules.

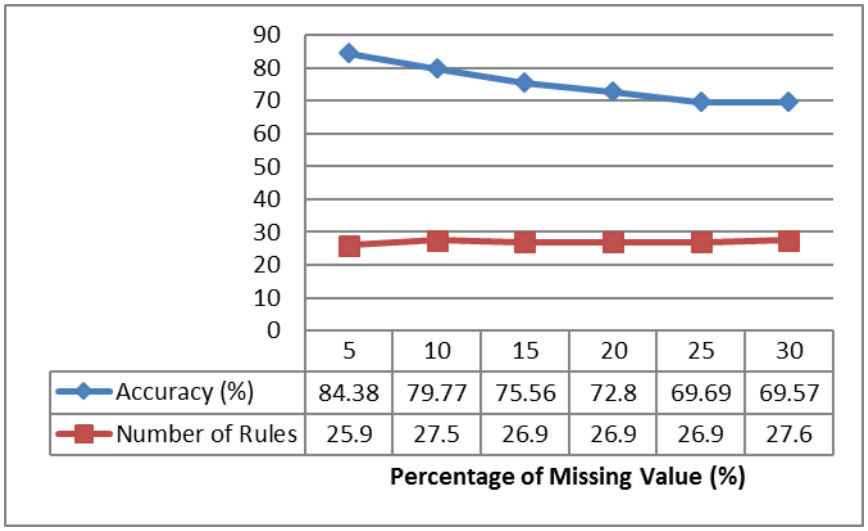

Figure 2 Graph of Comparison Number of Rules with Accuracy of Rules

In this study, the prediction results obtained strongly influenced by the distribution of decision classes in the training data. If the decision class distribution is uneven and there is a significant difference in frequency between the decision class with another, prediction results of the test data will tend to grade decision with the greatest frequency.

CONCLUSION

This paper has discussed regarding the classification algorithm to get the rules with rough set approach. Rules obtained by finding lower and upper approximation of incomplete dataset that has been converted into a complete dataset, which is replacing the missing value with the average value of the records that have the same decision value. After that, we obtained from certain rules of the lower approximation, while the possible rules obtained from the upper approximation. If rules have more specific and have a value less than or equal plausibility with other rules then will be removed.

The result of application of rough set algorithm on 1735 data contain 5%, 10%, 15%, 20%, 25%, and 30% missing value show that increasing missing value causes the accuracy of rules decreases, where accuracy is above 70% when percentage of missing value between 5% and 20%, but when the missing value increased to 25% and 30% accuracy is below 70%. This means that rules generated by rough set algorithm effectively to predict rainfall condition for dataset contain missing value less than 25%. Beside that the number of rules no affects the accuracy of rules.

REFERENCES

[1] Han, J., Kamber, M., & Pei, J., 2011, *Data Mining Concept and Techniques Third Edition*, Morgan Kauffman, USA.

[2] National Council of Applied Economic Research, 2010, *Impact Assessment and Economic Benefits of Weather and Marine Services*, (Online) (http://www.ncacr.org, accessed on September 8, 2014).

[3] Sadiq, A.T, Dualmi, M.G., & Shaker, A.S., 2013, "Data Missing Solution Using Rough Set Theory and Swarm Intelligence", *International Journal of Advanced Computer Science and Information Technology (IJACSIT)* Vol. 2, No. 3, 2013, Page: 1-16.

[4] Aprianti, W., and Mukhlash, I., 2014, The Application of Rough Set and Fuzzy Rough Set Based Algorithm to Classify Incomplete Meteorological Data, *Proceeding of 2014 International Conference on Data and Software Engineering*, Page: 177-182. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7062674.

[5] Chmielewski, M.R., Gryzmala-Busse, J.W., Peterson, N.W., and Than, S., 1993, The rule induction system LERS – A version for personal computers, *Foundations of Computing and Decision Sciences*, 18, Page: 181–212.

[6] Mukhlash, I., Iqbal, M., Astuti, H.M., and Sutikno, 2014, Performance Enhancement of CBS Algorithm Using FSGP and FEAT Algorithm, *Journal of Theoretical and Applied Information Technology*, Vol. 67 No.3.

[7] Wang, Q. and Fan, W., 2010, Cumulonimbus Forecasting Based on Rough Set and Artificial Immune Algorithm, *2010 Sixth International Conference on Natural Computation*, Page: 2856-2860.

[8] Shen, Q. and Jensen, R., 2007, Rough Sets, Their Extension and Applications, *International Journal of Automation and Computing* 04(3), Page: 217-228.

[9] Kryszkiewicz, M., 1998, Rough Set Approach to Incomplete Information Systems. *Information Science*, Vol . 112, No. 1, Page: 39-49.

[10] Hong, T., Wang, T., Wang, S., 2007, Mining Fuzzy β-Certain and β-Possible Rules from Quantitative Data Based on The Variable Precision Rough Set Model, *Expert System with Application* 32(1) (2007), Page: 223-232.

[11] Hong, T., Tseng, L., & Cien, B., 2009, Mining from Incomplete Quantitative by Fuzzy Rough Sets, *Expert Systems with Application*.