IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 9, Issue 3 (Mar. - Apr. 2013), PP 06-10
www.iosrjournals.org

A Comparative Study on Decision Rule Induction for incomplete data using Rough Set and Random Tree Approaches

M. Sandhya[1], Dr. A. Kangaiammal[2], Dr. C. Senthamarai[3],
[1](M.Phil. Scholar of Computer Science, Govt. Arts College (Autonomous), Salem -7, Periyar University, INDIA)
[2], [3] (Assistant Professor of Computer Applications, Govt. Arts College (Autonomous), Salem -7, Periyar University, INDIA)

Abstract: *Handling missing attribute values is the greatest challenging process in data analysis. There are so many approaches that can be adopted to handle the missing attributes. In this paper, a comparative analysis is made of an incomplete dataset for future prediction using rough set approach and random tree generation in data mining. The result of simple classification technique (using random tree classifier) is compared with the result of rough set attribute reduction performed based on Rule induction and decision tree. WEKA (Waikato Environment for Knowledge Analysis), a Data Mining tool and ROSE2 (Rough Set Data Explorer), a Rough Set approach tool have been used for the experiment. The result of the experiment shows that the random tree classification algorithm gives promising results with utmost accuracy and produces best decision rule using decision tree for the original incomplete data or with the missing attribute values (i.e. missing attributes are simply ignored). Whereas in rough set approach, the missing attribute values are filled with the most common values of that attribute domain. This paper brings out a conclusion that the missing data simply ignored yields best decision than filling some data in the place of missing attribute value.*

Keywords- *Random Tree, WEKA, ROSE2, Missing attribute, Incomplete dataset, Classification, Rule Induction, Decision Tree.*

I. Introduction

Data mining is the way of extracting useful information and discovering knowledge patterns that may be used for decision making [7]. Several data mining techniques are association rule, clustering, classification and prediction, neural networks, decision tree, etc. Application of data mining techniques concern to develop the methods that discover knowledge from data and then used to uncover the hidden or unknown information that is not apparent, but potentially useful [5]. Classification and Clustering are the important techniques in data mining. Classification groups data based on a classifier model while clustering groups the data based on the distance or similarity.

Rough set theory was introduced by Zdzisław Pawlak in early 1980's. The main aim of the rough set analysis is to find the approximation of concepts from the existing data. In order to deal with vagueness of data, rough set theory replaces every vague concept with two important concepts called the upper and lower approximation space. Lower approximation consists of those objects which are surely belong to the set and the upper approximation consist of those objects which do possibly belong to the set [13]. Rough set theory is basically used for finding:

a) Hidden patterns in data
b) Significance of attribute
c) Reduced subset of data
d) Dependency of attributes and so on.

There are various reasons as to why datasets are affected by missing attribute values. Sometimes the irrelevant values will not be recorded in the data as said by [4]. Another reason is that forgot to place the values in the table or mistakenly erased the data from the table. There are several approaches to handle the missing attribute values. The authors have found that filling the missing attribute with the most common attribute value is the worst method among all the approaches. Among all the nine approaches discussed, the two approaches namely, C4.5 and ignoring the missing attribute values are the best methods.

The paper is organized as follows: section 2 describes the related research work. Section 3 states the problem statement and section 4 describes the proposed method about filling the incomplete dataset by most common attribute value using rough set approach to generate rule induction and also about without filling the incomplete dataset to generate rule induction using Random Tree classification algorithm in data mining. Experimental results and performance evaluation are presented in section 5. Finally, section 6 concludes the work and points out some of the prospective future work.

II. Related Work

There are nine different approaches discussed for handling the missing attribute values. Ten input data files were used to apply and test the nine approaches for investigating the performance while handling missing attribute values. They fixed quality criterion for ten-fold cross validation as the error rate which needed to be average. The authors have concluded based on Wilcoxon matched-pairs signed rank test that the two approaches namely, C4.5 and ignoring the missing attribute values are the two best methods to handle the missing attribute values [4].

The scholars of [6] describe an ISOM-DM (Independent Self Organizing Maps) model that has been proposed for incomplete data handling in data mining. Compared with Mixture of Principal Component Analyzers (MPCA), mean method and standard SOM-based fuzzy map model, ISOM-DH model can be applied to more cases.

The work in [9] uses attribute value pair. These blocks are used to construct characteristic sets, characteristic relations, and lower and upper approximations for decision tables with missing attribute values. The authors in [9] conclude that an error rate for classification is smaller when missing attribute values are considered to be lost.

In [5] Characteristic relations are introduced to describe incompletely specified decision tables. For completely specified decision tables any characteristic relation is reduced to an indiscernibility relation. The basic rough set idea of lower and upper approximations for incompletely specified decision tables may be defined in a variety of different ways.

The work in [16] has made a comparative analysis of data mining classification technique and an integration of clustering and classification technique that helps in identifying large data sets. The integration of clustering and classification technique gives more accurate results than simple classification technique. It is also useful in developing rules when the data set is containing missing values. This integrated technique of clustering and classification gives a promising classification results with utmost accuracy rate.

III. Problem Statement

The problem here is to identify the best method of dealing with the missing attributes when decision making is important. This has been accomplished by comparing the results of rule generation through filling the incomplete dataset with most common attribute value using rough set approach and also in data mining without filling the incomplete dataset using Random Tree Classification for rule induction. Heart problem using 3-condition attributes and 1-ecision attribute with incomplete dataset has been considered for comparative study.

IV. Proposed Method

Rough set deals with vagueness and uncertainty of data. In rough set the incomplete dataset are described by their characteristic relation and complete decision tables are described by indiscernibility relations. Classification is the important technique in Data mining. Classification groups' data based on a classifier model. Using Random Tree classification algorithm decision is taken for incomplete dataset. Taking decision in the orginal table is the best method. Fig. 1 shows a general framework of a comparative analysis of two approaches for finding better rule induction for incomplete dataset. Fig. 2 shows the block diagram of steps of evaluation and comparison. Table 1 shows the incomplete dataset

In this experiment, decision attribute corresponds to heart problem and the condition attribute corresponds to blood pressure, chest pain and cholesterol. Apply two approaches for this table to generate rule in ROSE2 and WEKA tool. Apply classification technique in WEKA and attribute reduction in ROSE2.

In Rough set rule induction is based on the consistency and inconsistency of the table. Using reduct, the attribute is reducted and find the conisitency of the condition attribute and decision attribute. After reducing the attribute, with the consistancy of the table is chossen for the rule induction. Before finding the consistency and inconsistancy of the incomplete decision table, the table should be converted into a complete decison table. The rule is generated for the consistency of the table after removing a attribute.

In classification, the decision rule is generated using the decision tree. Random Tree considers a set of K randomly chosen attributes to split on at each node. Random tree, gives number of nodes by selecting all possible trees uniformly at random.

The complete descriptions of the incomplete dataset attribute value are presented in Table 1. In rough set, first the incomplete decision table is transformed to complete table by filling the missing attribute value using most common attribute value (i.e. the value of the attribute that occurs most often is selected as the value for all the missing values of the attribute). In WEKA, the decision rule is generated with the missing attribute value using Random tree classification algorithm. The complete decision table is shown in Table 2.

Table 1: Incomplete Decision Table

Case	Blood Pressure	Chest pain	Cholesterol	Heart Problem
1	High	?	High	Yes
2	?	Yes	?	Yes
3	?	No	?	No
4	High	?	High	Yes
5	?	Yes	Low	No
6	Normal	No	?	No

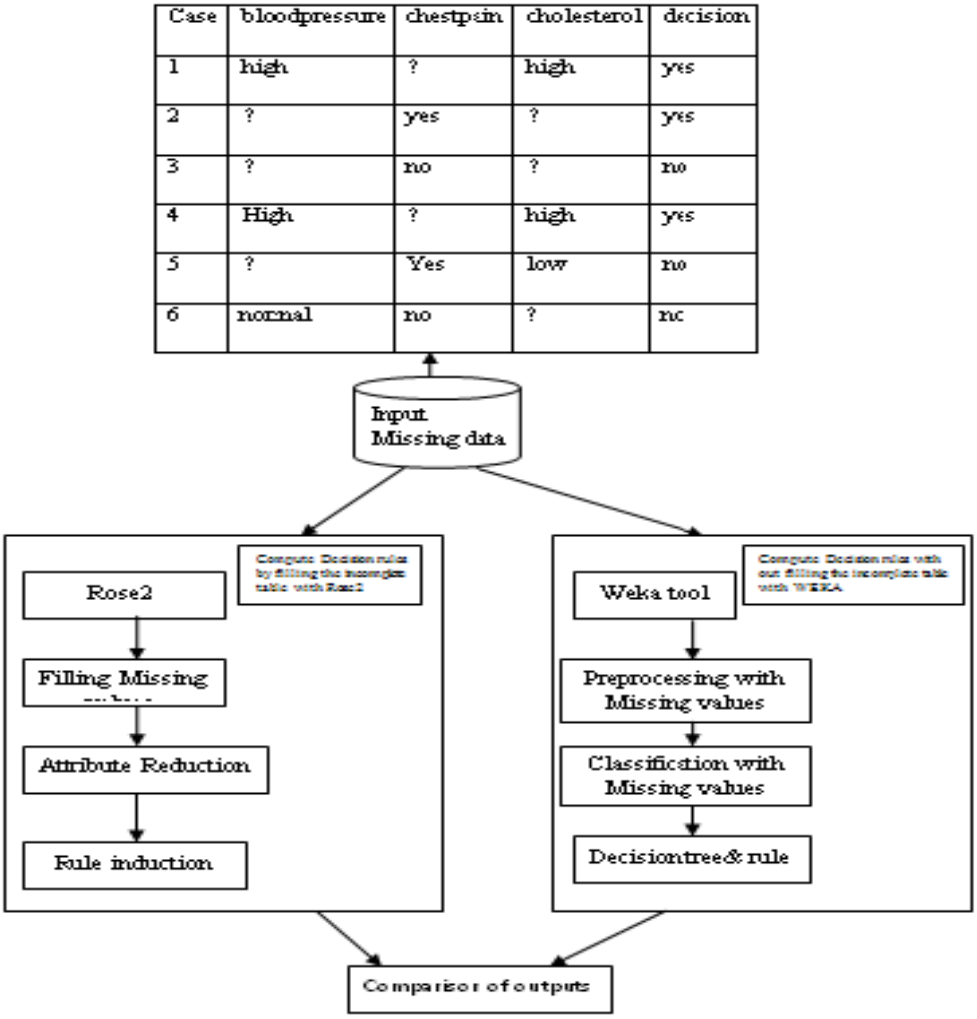

Figure 1: Architecture of the Proposed Work

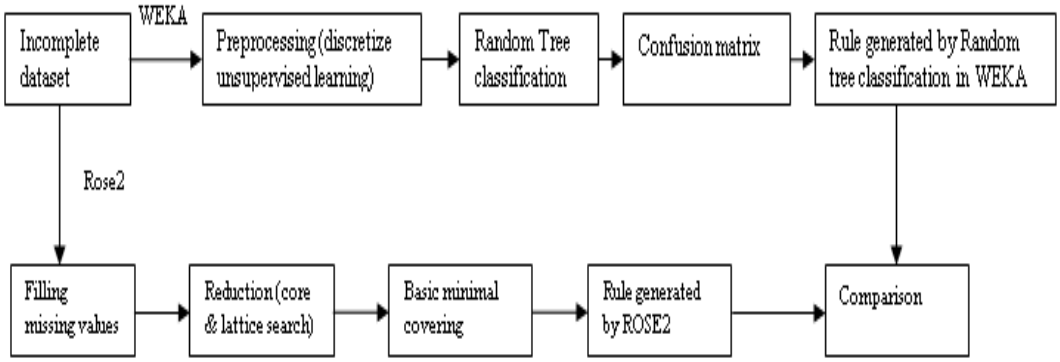

Figure 2: Block diagram of Evaluation and Comparison

V. Experiment Results And Performance Evaluation

In this experiment a comparative study of attribute reduction of rough set and classification technique of data mining technique for incomplete dataset on various parameters using missing attribute value in Heart problem data set containing 3 condition attribute and 1 decision attribute. During Rough set, the incomplete dataset is given as input to ROSE2 tool and fill the missing attribute value with the most common attribute value then reduce the attribute based on consistency and inconsistency of a table was implemented for rule generation. In data mining during simple classification, the training dataset is given as input to WEKA tool and the classification algorithm namely Random Tree was implemented.

The result of the experiment shows generating the rule in the original incomplete dataset produce best result than filling the attribute with most common attribute value. Table 2 shows the complete decision table generated by ROSE2 and figure3 shows the decision tree and rule generated by Random tree classification in WEKA.

Table 2: Complete Decision Table

ISF File Editor - [C:\Program Files\Roughset software\ROSE2\exar

File Attribute Example Options Window

Exit

	Blood_Pressure	Chest_Pain	Cholesterol	Heart_Problem [
1	high	yes	high	yes
2	high	yes	high	yes
3	high	no	high	no
4	high	yes	high	yes
5	high	yes	low	no
6	normal	no	high	no
*New				

Next is finding the Reduct and Core of the complete information table for generating rules based on Reduct of the complete information table. Reduct is a minimal subset of attributes that enables the relevancy and redundancy. A subset attribute is said to be relevant if it is predictive of the decision features, otherwise it is irrelevant. A subset attribute is considered to be redundant if it is highly correlated with other features. The result of analysis shows that there is only one reduct of Table 2. That is,

Reduct = {Chest Pain, Cholesterol}

The rule is generated after reducing the attribute by ROSE2 is

(Chestpain=yes) & (Cholesterol = yes) =>(Heartproblem=yes)
(Chestpain=no)=>(Heartproblem=no)
(Cholesterol=low)=>(Heartproblem=no)

In WEKA, the table1 is given as input. The decision rule is generated without filling the missing attribute value. With the original incomplete decision table the rule is generated. Using Random tree classification algorithm in WEKA the decision tree and rule is generated. The figure3 shows the decision tree and rule.

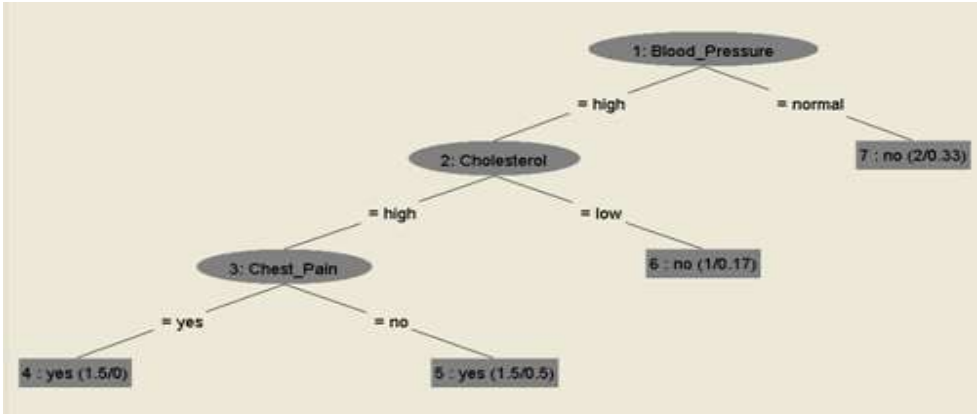

Figure 3: Decision Tree

Fig. 4 shows the rules framed by the random tree classification is as follows:
(Blood pressure = high & cholesterol = high & chest pain = high) = (heart problem = yes)
(Blood pressure = normal) = (heart problem = No)

(Blood pressure = high & cholesterol = low) = (heart problem = No)
(Blood pressure = high & cholesterol = high & chest pain = No) = (heart problem = yes)

```
RandomTree
==========

Blood_Pressure = high
|   Cholesterol = high
|   |   Chest_Pain = yes : yes (1.5/0)
|   |   Chest_Pain = no : yes (1.5/0.5)
|   Cholesterol = low : no (1/0.17)
Blood_Pressure = normal : no (2/0.33)

Size of the tree : 7

Time taken to build model: 0 seconds
```

Figure 4: Decision Rule

Observations and Analysis:

- It is observed that the random tree classification provides better rule than the reduction of attribute in rough set.
- Filling the most common attribute value for the incomplete decision is the worst method.
- Accuracy of RANDOM TREE classifier is high i.e. 100% (Table 3), which is highly required.

VI. Conclusion And Future Work

A comparative study of data mining classification and rough set attribute reduction for the incomplete dataset to generate the decision table has been performed. The presented experiment shows that the random tree classification algorithm is the best method for the rule generation of incomplete decision table. Because random tree generate the rule with the missing attribute values without filling the most common attribute value. Filling the missing attribute value with the most common attribute value is the worst method for the prediction. Therefore, it is observed that handling the incomplete decision table without filling the missing attribute value is best for prediction. In future, the work can be extended to use various other approaches for handling missing attribute values so as to observe the change in decision rules.

References

[1] Antony Rosewelt and Dr. Sampathkumar Vajeravelu, Mining Software Defects using Random tree, International Journal of Computer Science & Technology, IJCST Vol. 2, Issue 4, ISSN : 0976-8491(Online) | ISSN : 2229-4333(Print), Oct - Dec 2011.

[2] Frederick Livingston, Implementation of Breiman's Random Forest Machine Learning Algorithm, ECE591Q Machine Learning Journal Paper, Fall 2005.

[3] Grzymala-Busse, J.W. and Hu, M., A comparison of several approaches to missing attribute values in data mining. Proceedings of the Second International Conference on Rough Sets and Current Trends in Computing RSCTC-2000, Banff, Canada, October 16–19, 2000, 340–347.

[4] Grzymala-Busse, J.W., Rough set strategies to data with missing attribute values, Workshop Notes, Foundations and New Directions of Data Mining, The 3rd International Conference on Data Mining, Melbourne, FL, USA, November 19–22, 2003, 56–63.

[5] Hongyi Peng and Siming Zhu, Handling of incomplete data sets using ICA and SOM in data mining, Received: 2 September 2005 / Accepted: 24 April 2006 / Published online: 30 May 2006, Springer-Verlag London Limited 2006.

[6] http://idss.cs.put.poznan.pl/site/rose.html.

[7] J. Han and M. Kamber, Data Mining: Concepts and Techniques (Morgan Kaufmann, 2000).

[8] Jerzy W. Grzymala-Busse and Sachin Siddhaye, Rough Set Approaches to Rule Induction from Incomplete Data, Proceedings of the IPMU'2004, the 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Perugia, Italy, July 4–9, 2004, vol. 2, 923–930.

[9] Kash Barker et al, Learning From Student Data, Proceedings of the 2004 Systems and Information Engineering Design Symposium, Mathew H. Jones, Stephen D. Patek, and Barbara E. Towney eds. 2004. pp79-86.

[10] Marzena Kryszkiewicz, Rules in incomplete Information Systems, Information Science 113(1999), 271-292.

[11] Nambiraj Suguna and Keppana Gowder Thanushkodi, Predicting Missing Attribute Values Using k-Means Clustering, Journal of Computer Science 7 (2): 216-224, 2011

[12] Pawlak, Z, Rough Set, International Journal of Computer and Information Sciences (1982) 341–356.

[13] Pawlak, Z, Rough Sets. Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht, Boston, London (1991).

[14] Renu Vashist, M.L Garg, A Rough Set Approach for Generation and Validation of Rules for Missing Attribute Values of a Data Set, International Journal of Computer Applications (0975 – 8887) Volume 42– No.14, March 2012.

[15] Varun Kumar, Nisha Rathee , Knowledge discovery from database Using an integration of clustering and classification", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No.3, March 2011.

[16] Weka 3- Data Mining with open source machine learning software available from: - http://www.cs.waikato. ac.nz/ml/ weka.

[17] Z. Pawlak, Rough Sets and Intelligent Data Analysis [J]. Information Sciences, 2002, 147(1-4) 1-12.

[18] Zdzisław Pawlak, Rough set theory and its applications, journal of telecommunication and technologies, 2002.