A Rough Set Approach for Generation and Validation of Rules for Missing Attribute Values of a Data Set

Renu Vashist
School of Computer Science and Engineering
Shri Mata Vaishno Devi University,
Katra, (J & K), India.

M.L Garg
School of Computer Science and Engineering
Shri Mata Vaishno Devi University,
Katra, (J & K), India.

ABSTRACT
Data mining has emerged as most significant and continuously evolving field of research because of it's ever growing and far reaching applications into various areas such as medical, military, financial markets, banking etc. One of the most useful applications of data mining is extracting significant and earlier unknown knowledge from real-world databases. This knowledge may be in the form of rules. 'Rule generation' from raw data is a very effective and most widely used tool of data mining. Real life data are frequently imperfect, erroneous, incomplete, uncertain and vague. There are so many approaches for handling missing attribute values. In this paper we use the most common attribute value approach i.e. replacing all the missing attribute values by most frequently occurring attribute value and thereby completing the information table. Subsequently, we find the reduct and core of the complete decision table and verify that the reduct and core find by our method is same as the reduct and core find by ROSE2 software. Thereafter we generate the rules based on reduct. Our results are validated by conducting the same rough set analysis on the incomplete information system using the software ROSE2.

General Terms
Artificial Intelligence, Data mining, Rough set, Decision tables.

Keywords
Data mining, Knowledge discovery from database, Machine learning, Reduct, Core, Missing attribute values, Rule generation.

1. INTRODUCTION
Knowledge Discovery in Databases (KDD) is a process comprises of many stages. One of the important stage of KDD is usually Data Mining, which is the sequence of operations that leads to discovery of new, interesting, useful and non-trivial patterns from data[17]. Data mining is basically analyzing the data from different perspective and then finding the correlation or previously unknown patterns from data [18][19]. The most important data mining technique which search through the entire dataset is association rule generators that finds the rules revealing the nature and frequency of relationships between data entities. Rough set can be used as a tool of data mining that is used for rule generation. Pawlak[11] introduced rough sets theory in the early 1980's. It can be seen as a new mathematical approach to deal with vagueness and uncertainty [11]. Rough set theory is relatively a new decision making tool which can be applied to wide range of fields such as medicine, banking, financial markets, military and education etc. The rough set philosophy is founded on the assumption that with every object of the universe of discourse there is some information associated (data, knowledge). We assume that the input data files are in the form of a table known as a *decision table or information table*. In this table, each column represents one *attribute* and each attribute represents some feature of the examples, and each row represents an *example* by all its attribute values. There are basically two types of attributes in a decision table one is called as condition attributes and other is called as decision attribute. Condition attributes are called independent variables of the decision table and decision attribute is called dependent variable[1]. In Most of the cases each decision table has only one decision attribute, there may be any number of condition attributes in the decision table.

Objects characterized by the same information are indiscernible (similar) in view of the available information about them. The indiscernibility relation generated in this way is the mathematical basis of rough set theory[12]. In order to deal with vagueness, rough set theory replace every vague concept with two precise concept called the upper and lower approximation. Lower approximation consist of those objects which surely belongs to the set where as upper approximation consist of those objects which possibly belongs to the set[1][11]. The elements which are in the lower approximation are surely belong to the set whereas the elements which are in the upper approximation may or may not belong to the set. Moreover the elements which are present in upper approximation and not present in the lower approximation gives us boundary region cases[2]. A non empty boundary region represents the rough set[12]. 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.

However, in real life applications, input data presented in the form of table is usually missing or in other word we can say that decision tables are incompletely specified and some attribute values are frequently absent. Handling missing attribute values in rough set theory is a big challenge. The concept of rule induction from incomplete data set is first given by Jerzy W. Grzymala-Busse. There are basically two main reasons for the attribute values to be missing; either they were "lost" means originally the attribute value was known and due to some unknown reason it is erased and the "do not care condition" the original values were not recorded since they were irrelevant[4]. The first rough set approach to missing attribute values, when all missing values are lost, was described in 1997 in [8]. On the other hand, decision tables in which all missing attribute values are " do not care" condition were described for the first time in [4].

In general, incomplete decision tables are described by characteristic relations, in a similar way as complete decision tables are described by indiscernibility relations[5][6]. In rough set theory, for complete decision tables, once the indiscernibility relation is fixed and the concept (a set of cases) is given, the lower and upper approximations are unique. But for incomplete decision tables, a given characteristic relation and concept, there are three different possibilities to define lower and upper approximations, called singleton, subset, and concept approximations [6][7]. Singleton lower and upper approximations were studied in [9][10][13][14][15]. It is observed in[6] that singleton lower and upper approximation are not applicable in data mining. Where as rules generated from concept lower and upper approximation are most significant.

2. INCOMPLETE DECISION TABLE

In our information table all condition attribute values are not specified, i.e., they are missing or we can say that the decision table is incomplete. When all the condition attribute values are not specified in the table then that table is known as incomplete decision table. We know that all missing attribute values are denoted either by "?" or by "*". The lost values will be denoted by "?", and "do not care" conditions will be denoted by "*". For each case at least one attribute value is specified"[6]. For this particular case we assume that some of the condition attribute values are lost and all decision attribute values are present, i.e., they are not missing. In table 1 the lost condition attribute values are donated by "?".

Table1 An incomplete Decision Table

Cases	Condition Attributes			Decision
	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

In the Decision Table 1 there are three condition attributes and one decision attribute.

Condition attributes = {Blood Pressure, Chest Pain, Cholesterol}

and

Decision attribute = {Heart Problem}

Here, the existence of Heart Problem is dependent on values of Blood Pressure, Chest Pain, and Cholesterol.

Table 2 Values of all Attributes

	Attributes	**Nominal Values**
Condition Attributes	Blood Pressure, Chest Pain, Cholesterol	High, Normal Yes, No High, Low
Decision Attribute	Heart Problem	Yes, No

This paper adopts following four step methodology to validate reduct and rules.

Step 1 Completing the incomplete decision table with the help of the method of most common attribute value i.e replacing every missing attribute value by most frequently occurring attribute value.

Step2 Finding the Reduct and Core of the complete information table and validating that reduct and core find by our method is same as ROSE2 software.

.Step 3 Generating Rules based on Reduct of the complete information table.

Step 4 Validating Rules with the help of ROSE2 software.

2.1 Preprocessing of Incomplete Information Table

There are many approaches of completing the missing attribute values such as most common attribute value, concept most common attribute value, method of assigning all possible values of attribute, event covering method, C4.5, a special LEM2 algorithm[6]. This paper deals with completing the information table with the help of *Most Common Attribute Value.* It is one of the simplest methods to deal with missing attribute values. The CN2 algorithm uses this idea[3]. The value of the attribute that occurs most often is selected as the value for all the missing values of the attribute.

In decision table1 replacing the missing attribute value of Blood Pressure by attribute value 'high' since the value 'high' appears more frequently then the value 'low'. Similarly for attribute Chest Pain either value 'Yes' or value 'no' can be used since both of them occur twice. We randomly choose value 'yes' for Chest pain. For Cholesterol attribute the value 'high' occurs more often.

Table3 is obtained after substituting 'high' in place of '?' for Blood pressure attribute. Similarly replacing the value 'yes' by '?' for Chest Pain attribute and value 'high' by '?' for Cholesterol in Table1. After substituting these values in table1 we get a complete decision table.

Table 3 Complete Decision Table

Cases	**Blood Pressure**	**Chest Pain**	**Cholesterol**	**Heart Problem(D)**
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

2.2 Finding Reduct and Core of Complete Information Table

One of the most important contributions of rough set theory to the data analysis field is that it can remove superfluous information. Removal of duplicate data from the information system is one of the central concept in rough set theory. The concepts which perform this task are reduct and core. Reduct is a set of attributes that preserves partition[2]. It means that a reduct is a minimal subset of attributes that enables the same classification for elements of the universe as the whole set of attributes. In other words, attributes that do not belong to a reduct are superfluous with regard to classification of elements of the universe. The core is the set of all indispensable attributes of the set. The core contains the most significant attributes of the set. We cannot remove any of the elements from the core without losing information from the set. The following is an important property that connects the notion of the core and reducts[1]

Core = $\cap$ Red(A)

where Red(A) is the set of all reducts of set.

Now using a simultaneous method of finding the reduct and core.

For finding reduct and core we must understand the concept of consistency and inconsistency. A table is known as consistent if all the same values of condition attribute leads to the same decision attribute. But if for the same values of condition attribute we have different decision attribute values then that table is known as inconsistent.
After removing any one of the attribute from the table we will check the consistency of the remaining table. If remaining table is consistent than the attributes of the table form a reduct set. But if there is inconsistency in the remaining table then the attribute which has been removed from the table gives us the core attribute. For finding the reduct and core, remove attribute Blood Pressure from Table 3 and we get a new table i.e Table4

Table 4

Cases	Chest Pain	Cholesterol	Heart Problem(D)
1	Yes	High	Yes
2	Yes	High	Yes
3	No	High	No
4	Yes	High	Yes
5	Yes	Low	No
6	No	High	No

After interpreting and analyzing Table4 we find that there is no inconsistency in the table. Hence,

Reduct = {Chest Pain, Cholesterol}

Similarly after removing attribute Chest Pain from Table 3 we get a new table i.e Table5.

Table5

Cases	Blood Pressure	Cholesterol	Heart Problem(D)
1	High	High	Yes
2	High	High	Yes
3	High	High	No
4	High	High	Yes
5	High	Low	No
6	Normal	High	No

An introspection of table5 indicates that there is inconsistency in the table5. All the condition attribute values for cases 1,2,3 and 4 are same but the decision value for case3 is different then that for cases 1, 2 and 4. So there is inconsistency in the table. Hence Chest Pain is the core of the decision table 3.

Core1 = {Chest Pain}

Similarly after removing attribute cholesterol from Table3 we get Table 6.

Table6

Cases	Blood Pressure	Chest Pain	Heart Problem(D)
1	High	Yes	Yes
2	High	Yes	Yes
3	High	No	No
4	High	Yes	Yes
5	High	Yes	No
6	Normal	No	No

On studying the table6 inconsistency in the table is reveled. All the condition attribute values for cases 1,2, 4 and 5 are same but the decision value for case5 is different form cases 1, 2 and 4 so there is inconsistency in the table. Hence Cholesterol is the core of the decision table 3.
Core2 = {Cholesterol}

The result of our analysis shows that there is only one reduct and two core of Table3.

Reduct = {Chest Pain, Cholesterol}

The reduct find by ROSE2 software is same as shown in Fig1.

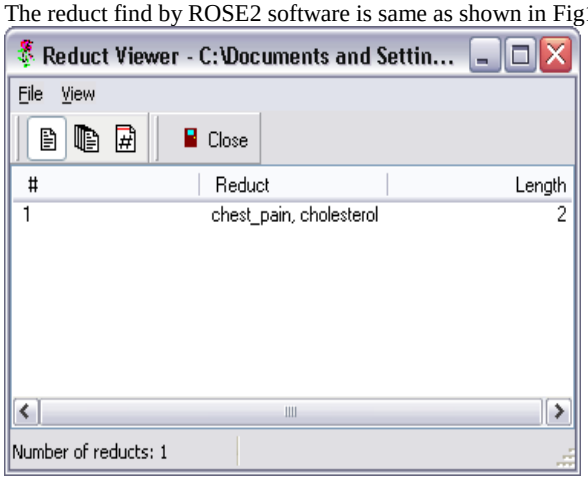

Fig 1

and
Core 1 = {Chest Pain} and
Core 2 = {Cholesterol}
The core find by ROSE2 software is shown in Fig2
We can validate that reduct and core find by our method is same as reduct and core find by ROSE2 software.

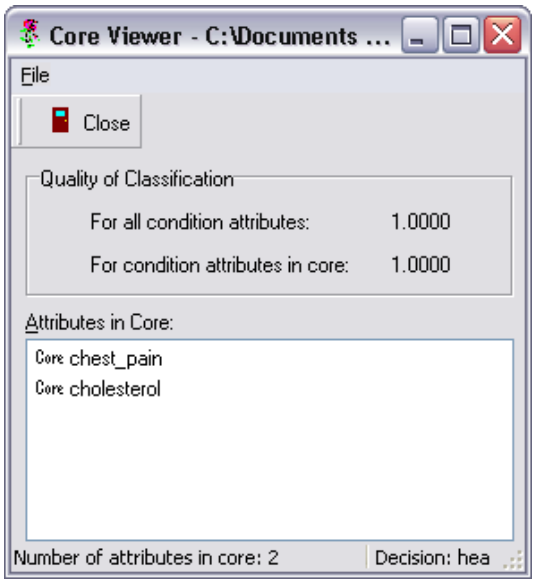

Fig 2

2.3 Generations of Rules

In this step we generate rules based on the reduct of the decision table. We use table4 i.e reduct for generating rules.
Reduce this table for generating the rules. In table 4 cases 1, 2 and 4 are similar i.e all the condition and decision attribute values are same that is these three rows are duplicate. We can remove any two of them without affecting the information in the

decision table. We remove cases 2 and 4. Similarly cases 3 and 6 are same, so case 6 is removed from the table4. After removing cases 2, 4, 6 from the table we are left with a new table i.e Table7

Table 7

Cases	Chest Pain	Cholesterol	Heart Problem(D)
1	Yes	High	Yes
3	No	High	No
5	Yes	Low	No

Decision rules are basically implication such as φ→ψ. If φ occur then ψ also occur, where φ and ψ are the condition and decision attribute respectively[2]. Decision rules are used to measure the closeness of concepts.

Rule 1. (Chest_pain = Yes) & (Cholesterol = High) => (Heart_Problem = Yes)

Rule 2. (Chest_pain = No) & (Cholesterol = High) => (Heart_Problem = No)

Rule 3. (Chest_pain = Yes) & (Cholesterol = Low) => (Heart_Problem = No)

These rules can be used for taking important decisions.

2.4 Validating Rules with the help of Rose2 Software

The rough set analysis was conducted on the incomplete information system, using the software ROSE2 which was created at the Laboratory of Intelligent Decision Support systems of the Institute of Computing Science in Poznan [16].

Preprocessing of the incomplete decision table is done using missing values in the ROSE2 software.

The analysis of the transformed information system shows the following results. The rules generated by us is same as rules generated by ROSE2 software. We validate that ROSE2 software also use most occurring attribute value for filling the missing attribute value. Rules generated by ROSE2 software is shown in Fig3.

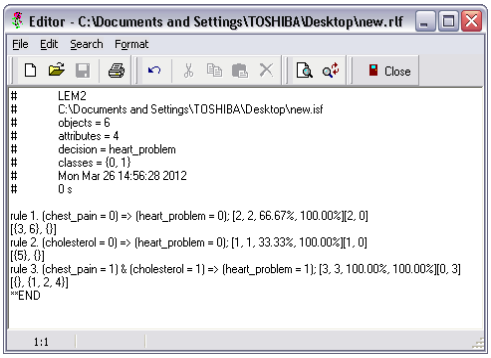

Fig 3

We validate that the number of rules generated by our method is same as the rules generated by Rose2 software.

3. CONCLUSION

There are so many methods for filling the missing attribute values but for this particular paper we use most common attribute value that is replacing each missing attribute value by most frequently occurring value. Rule generation is an important aspect of data mining provided the number of the rules generated are minimum and at the same time most representative of the underlying data set. In this paper a unique approach has been adopted to produce the reduct and core of the decision table and these reduct and core are validated using ROSE2 software. Finally decision rules are generated using the reduct of the decision table. Thereafter these rules are also validated using ROSE2 software.

4. REFERENCES

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

[2] Z Pawlak, Andrzej Skowron. Rudiments of rough sets[J]. Information Sciences, 177(2007) 3-27.

[3] Clark, P. Niblett, T.: The CN2 induction algorithm. Machine Learning **3** (1989) 261–283.

[4] Grzymala-Busse, J. W.: On the unknown attribute values in learning from examples. Proc. of the ISMIS-91, 6th International Symposium on Methodologies for Intelligent Systems, Charlotte, North Carolina, October 16–19, 1991, Lecture Notes in Artificial Intelligence, vol. 542. Springer-Verlag, Berlin Heidelberg New York (1991) 368–377.

[5] Grzymala-Busse, J. W.: LERS—A System for Learning from Examples Based on Rough Sets. In: Slowinski, R. (ed.): *Intelligent Decision Support. Handbook of Applications and Advances of the Rough Sets Theory*. Kluwer Academic Publishers, Boston MA (1992) 3–18.

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

[7]. 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.

[8]. Grzymala-Busse, J.W. and A. Y. Wang A.Y.: Modified algorithms LEM1 and LEM2 for rule induction from data with missing attribute values. Proc. of the Fifth International Workshop on Rough Sets and Soft Computing (RSSC'97) at the Third Joint Conference on Information Sciences (JCIS'97), Research Triangle Park, NC, March 2–5, 1997, 69–72.

[9] Kryszkiewicz, M.: Rough set approach to incomplete information systems. Proceedings of the Second Annual Joint Conference on Information Sciences, Wrightsville Beach, NC, September 28–October 1, 1995, 194–197.

[10].Kryszkiewicz, M.: Rules in incomplete information systems. *Information Sciences* 113 (1999) 271–292.

[11]. Pawlak, Z.: Rough Sets. *International Journal of Computer and Information Sciences* (1982) 341–356.

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

[13]. Stefanowski, J.: Algorithms of Decision Rule Induction in Data Mining. Poznan University of Technology Press, Poznan, Poland (2001).

[14.] Stefanowski, J. and Tsoukias, A.: On the extension of rough sets under incomplete information. Proceedings of the 7th International Workshop on New Directions in Rough Sets, Data Mining, and Granular-Soft Computing, RSFDGrC'1999, Ube,Yamaguchi, Japan, November 8–10, 1999, 73–81.

[15] Stefanowski, J. and Tsoukias, A.: Incomplete information tables and rough classification. *Computationa l Intelligence* 17 (2001) 545–566.

[16] Prędki B, Wilk S (1999) Rough set based data exploration using ROSE system. In: Ras ZW, Skowron A (Eds.), Foundations of Intelligent, Lecture Notes in Artificial Intelligence, vol. 1609, Springer, Berlin, 172-180

[17]Jiawei Han, and Micheline Kamber, *Data Mining: Concepts and Techniques*. California Morgan Kaufmann Publishers, 2000.

[18] R. Brachman, T. Khabaza, W. Kloesgen, G. Piatetsky-Shapiro, and E. Simoudis, Industrial Applications of Data Mining and Knowledge Discovery, *Communzcatzons of ACM*, vol. 39, no. 11.1996.

[19] Communications of The ACM, special issue on Data Mining, vol. 39, no. 11.