Volume 5, Issue 12, December 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

RST based Dataset Reduction using Decision Tree

[1]P. N. Senthil Prakash*, [2]S. Muthu Kumar, [3]S. Manoj Kumar
[1, 3] AP(Sr.G)/CSE, KPR Institute of Engineering & Technology, Coimbatore, India
[2]AP/IT, MEPCO Engineering College, Sivakasi, India

Abstract*: Rough set theory has evolved as one of the most important technique used for feature selection as a result of contemporary developments in data mining. One of the cardinal uses of Rough set theory is its application for rule generation. More often attribute reduction poses a major challenge for developing the theory and applications of rough set. This paper proposes a unique mathematical approach for determining the most important attribute and to construct the decision tree. The proposed algorithm firstly reduces the large volume of dataset which contain redundant instance. These redundant instance doesn't make any contribution to take decision hence can be deleted from the dataset. After reducing the volume of the dataset decision tree is constructed through rough set. The main concept of rough set theory is degree of dependency which is used in the proposed algorithm to select splitting attribute on the compressed data. Thus the proposed algorithm reduces the complexity of tree and in addition increases the accuracy. We have used some UCI machine learning repository (fertility dataset). Since only the subset of the features is included in the decision tree, the proposed algorithm gives better performance.

Keywords: Data mining, Rough set, Dataset reduction, Decision tree, ID3.

I. INTRODUCTION

In the digital era, data stored in a database and which is used for applications is huge. This explosive growth in data and databases has generated an urgent need for new techniques and tools that can intelligently and automatically transform the processed data into useful information and knowledge. Hence data mining has become a research area with increasing importance. Classification in data mining has gained a lot of
importance in literature and it has a great deal of application areas from medicine to astronomy, from banking to text classification. The aim of the classification is to find the similar data items which belong to the same class. For example, as the tail length, ear size, number of teeth etc are the variables which may vary from one specie to another, the variables 'cat' and 'dog' will be determined according to the values of the other variables. Classification is a predictive model of data mining that predicts the class of a dataset item using the values of some other variables. Many algorithms have been introduced with different models. Decision tree is widely used classification algorithm which can be useful for finding structures in high dimensional spaces and in problems with mixed data, continuous and categorical. ID3, C4.5 and CART algorithms are best known decision tree algorithms. The most commonly used ID3 algorithm uses information gain to measure impurity of the data. It selects condition attribute as splitting attribute which have highest information gain. The ID3 algorithm is simpler algorithm but it have strong disadvantage that it selects those condition attribute as a splitting attribute which have different attribute values whether this attribute contain noise or irrelevant information.

II. BASIC CONCEPT OF ROUGH SET

2.1 Theory

Rough set theory, introduced by Zdzislaw Pawlak in the early 1980s is a new mathematical tool to deal with vagueness and uncertainty [2][3][4]. Rough set theory can be seen as a mathematical approach to intelligent data analysis and data mining.
Suppose, we are given an information system
$S = (U , A),$ $X \subseteq U$ and $P \subseteq A$
where U and A, are finite, nonempty sets and called as the universe, and the set of attributes, respectively. Set A will contain two disjoint sets of attributes, called condition and decision attributes and the system is donated by $S = (U , C , D)$ where C is called condition attribute and D is called decision attribute. With every attribute $a \in A$ we associate a set V_a, of its values, called the domain of a.
Now we define two approximations $\underline{P (X)}$ and $\overline{P (X)}$ called the P-lower and the P-upper approximation of X respectively where

$$\underline{P (X)} = \bigcup x \in U \{P (X) : P (X) \subseteq X \}$$

and

$$\overline{P (X)} = \bigcup x \in U \{P (x) : P (x) \cap X \neq \varphi \}$$

Lower approximation will consist of all the members which surely belongs to the set and Upper approximation consist of all the members which possibly belongs to the set. The boundary region is given by the set difference $P(X) - P (X)$ consists of those objects that can neither be ruled in nor ruled out as members of the target set X. If the boundary region

is empty i.e $\underline{P(X)} = \overline{P(X)}$ then the set is crisp otherwise the set is rough[2]. Rough set theory can determine whether there is any redundant information in the data and if it is there, then can we find essential data required for our applications. The accuracy of the approximation to the set X from the elementary subsets is measured as the ratio of the lower and the upper approximation size. The ratio is equal to 1, if no boundary region exists, which indicates a perfect classification. In this case, deterministic rules for the data classification can be generated. Thus, a set X with accuracy equal to 1 is crisp, otherwise X is rough [3].

2.2 Reduct and Core

Reduct and core are the two most important concept of rough set theory. Reduct is a reduced subset of original set which retains the accuracy of original set. Reduct is often used in the attribute selection process to reduce unnecessary attributes towards decision making applications [1].

1. Reduct of a decision table is a set of condition attributes that is sufficient to define the decision attribute.
2. Any reduct enables us to reduce condition attribute.
3. A reduct does not contain redundant attribute toward a classification task.
4. It reduces computation cost for rule generation.
5. Finding all the reduct is an NP-hard problem.

In a decision table we may find multiple reduct and some rule would appear more frequently in some reduct than others. There are so many methods of finding reduct of a decision table. The reducts can be obtained by using the reduct generation algorithms. Using the discernibility matrix, the reduct of a decision table can be found[1]. The core can be found as the set of all singleton entries in the discernibility matrix. The reduct is the minimal element in the discernibility matrix, which intersects all the element of the discernibility matrix.

2.3 Reduct and Core

Reduct and core are the two most important concept of rough set theory. Reduct is a reduced subset of original set which retains the accuracy of original set. Reduct is often used in the attribute selection process to reduce unnecessary attributes towards decision making applications[1].

1. Reduct of a decision table is a set of condition attributes that is sufficient to define the decision attribute.
2. Any reduct enables us to reduce condition attribute.
3. A reduct does not contain redundant attribute toward a classification task.
4. It reduces computation cost for rule generation.
5. Finding all the reduct is an NP-hard problem.

In a decision table we may find multiple reduct and some rule would appear more frequently in some reduct than others. There are so many methods of finding reduct of a decision table. The reducts can be obtained by using the reduct generation algorithms. Using the discernibility matrix, the reduct of a decision table can be found[1]. The core can be found as the set of all singleton entries in the discernibility matrix. The reduct is the minimal element in the discernibility matrix, which intersects all the element of the discernibility matrix.

III. DATA REDUCTION AND FINDING THE RULES

Rough sets have many applications in the field of Knowledge Discovery in Databases (KDD), such as feature selection, data reduction, discretization, etc. When a dataset contains irrelevant (dispensable) features the same may be eliminated and thereby reducing the dimension of the problem. Thereafter, Rough sets can be used to find subsets of relevant (indispensable) features [8].

The volume of data is increasing day by day. In many real applications, it is very difficult to find which attributes are important for a particular task and which attributes are not so important. Hence identifying the relevant features is important for the reduction of the volume of data. The aim of data reduction is to find a minimal subset of relevant attributes that have all the essential information of the data set, thus the minimal subset of the attributes can be used instead of the entire attributes set for rule discovery.

3.1 Decision Table

Rough set theory can be considered as an extension of classical set theory. The basic concept of the RST is the notion of approximation space, that is with every object of universe we associate some information i.e. Data and Knowledge. Every example of the Rough set is organized in the form of information table, whose columns are labeled as condition and decision attributes and rows of the table contain the example[3]. Entries in the table represents the attribute values.

Table 1 is a decision table whose decision attribute is D and condition attributes are $\{x, y, z, w\}$.

Table 1 Decision table

	season	age	fever	smoking	diagnosis
1	-	0.69	0	0	N

	0.33				
2	-0.33	0.67	0	0	N
3	-0.33	0.67	0	-1	0
4	-0.33	0.94	0	1	0
5	1	0.69	-1	-1	0
6	1	0.67	0	-1	N
7	1	0.67	-1	1	0
8	1	0.67	-1	0	0
9	-1	0.53	1	-1	N
10	-1	0.53	1	-1	N

From Table 1 it is easy to see that for example 9 and 10 all the values of the condition attributes are same except for the values of decision attributes. We can say that Table1 is inconsistent because example 9 and 10 are conflicting (or are inconsistent) for both examples the value of all condition attribute is the same, yet the decision value is different.

3.2 Lower and Upper Approximations

Rough set theory offers a tool to deal with inconsistencies[4]. For each concept *X* the greatest definable set contained in *X* and the least definable set containing *X* are computed. The former set is called a *lower approximation* of *X* the latter is called an *upper approximation* of *X*. In the case of Table 1, the elementary sets are {1},{2},{3},{4},{5},{6},{7},{8},{9,10}

Now, let us consider the concept for the table1. We can define decision attributes and elementary set associated with the decision- as subset of the set of all examples with the same value of decision. Such subset are called concept. There are three concepts in Table1.

A1 = {1,2,6} for decision 1

A2 = {3,4,5,7,8} for decision 2

We can easily find lower and upper approximation of these three concepts.

Lower approximation is

$P(X) = \{1,2,3,4,5,6,7,8\}$

Upper approximation is

$P(X) = \{1,2,3,4,5,6,7,8,9,10\}$

The Boundary region is

Upper Approximation – Lower approximation = {9, 10}

We are again drawing the consistent part of Table1. By removing conflicting example i.e. 9 and 10.

Table2 Consistent Part of Table1

	season	**age**	**fever**	**smoking**	**diagnosis**
1	-0.33	0.69	0	0	N
2	-0.33	0.67	0	0	N
3	-0.33	0.67	0	-1	0
4	-0.33	0.94	0	1	0
5	1	0.69	-1	-1	0
6	1	0.67	0	-1	N
7	1	0.67	-1	1	0

8	1	0.67	-1	0	0

3.3 Rule Generation

Now we will generate the rules based on reduct and core of Table 2. Reduct is the reduced set of relation that conserves the same inductive classification of Relation. The set *P* of attributes is the reduct (or covering) of another set *Q* of attributes if *P* is minimal and the indiscernibility relations, defined by *P* and Q are same.

Core = ∩ reduct

Reduct of table2 are {season, age, smoking}, {season, age, fever} and core of the table2 is attribute {season, smoking}. We can't eliminate attribute w because this is the most important attribute of the Table2. By using the confidence or strength (α) we will find another indispensible attribute of the table. The confidence or strength for an association rule *season→ diagnosis* is the ratio of number of example that contains *season* U *diagnosis* to the number of example that contain *season*.

For Table 2 we can calculate the strength of attribute season, smoking and fever as follows:

We can find the strength of rules for attribute *season*

(*season* = -0.33) → (*diagnosis* =N) strength of this particular rule comes out to be 50%.
(*season* = -0.33) → (*diagnosis* =O) strength of this particular rule comes out to be 50%.
(*season* = 1) → (*diagnosis* =N) strength of this particular rule comes out to be 33%.
(*season* = 1) → (*diagnosis* =O) strength of this particular rule comes out to be 75%.
(*season* = -1) → (*diagnosis* =N) strength of this particular rule comes out to be 100%.
(*season* = -1) → (*diagnosis* =O) strength of this particular rule comes out to be 0%.

Similarly we can find the strength of rules for attribute *smoking*

(*smoking* = 0) → (*diagnosis* =N) strength of this particular rule comes out to be 75%.
(*smoking* = 0) → (*diagnosis* =O) strength of this particular rule comes out to be 33%.
(*smoking* = 1) → (*diagnosis* =O) strength of this particular rule comes out to be 100%.
(*smoking* = -1) → (*diagnosis* =O) strength of this particular rule comes out to be 75%.

and

(*age*= 0.69) → (*diagnosis* =N) strength of this particular rule comes out to be 100%.
(*age* = 0.67) → (*diagnosis* =N) strength of this particular rule comes out to be 40%.
(*age* = 0.67) → (*diagnosis* =O) strength of this particular rule comes out to be 60%.
(*age* = 0.94) → (*diagnosis* =N) strength of this particular rule comes out to be 100%.

From these calculations we can easily find that attribute *smoking* is indispensible among other attributes because the strength of rules for attribute *smoking* is maximum. The reduct of the set {*season*, *smoking*, *age*, *fever*} is { *season*, *smoking* }. Table2 can be reduced to Table 3 as follows.

Table 3

	season	**smoking**	**diagnosis**
1	-0.33	0	N
2	-0.33	0	N
3	-0.33	-1	O
4	-0.33	1	O
5	1	-1	O
6	1	-1	N
7	1	1	O
8	1	0	O

Reduce Table3 by eliminating the same values of decision and condition attributes i.e we can merge different rows that has the same values for condition and decision attributes. This method is called Row Reduction

Table 4

	season	**smoking**	**diagnosis**
1	-0.33	0	N
2	-0.33	-1	O
3	1	-1	O
4	1	-1	N
5	1	1	O

Find out the core of each example

We will find the core of the Table 4 in such a manner that the table will remain consistent. If we eliminate w =A there are two decision values 1 and 2. It means that based on attribute w we cannot make a unique decision, thus the value of y cannot be eliminated. Similarly if we eliminate y = R there are two decision values 2 and 3 It means that based on attribute y we cannot make a unique decision, thus the value of w cannot be eliminated. Now table 4 becomes

Table 5

	season	**smoking**	**diagnosis**
1	-0.33	0	N
2	-0.33	-1	O
3	1	-1	O
4	1	-1	N
5	1	1	O

Table 5 shows the core of each example. We can further reduced Table 5 by merging duplicate rows. Now we again eliminate the identical rows.

Table 6

	season	**smoking**	**diagnosis**
1	*	O	N
2	-0.33	-1	O
3	1	1	O

Now, no further reduction is possible. Table 6 gives us the decision rules. Followings are the decisions rules based on reducts:

IF *season* → -0.33 AND *smoking* →0 THEN *diagnosis* → N

IV. CONCLUSION

This paper presents an approach for determining the most important attribute on the basis of strength of an association. It is one of the most promising and new analytical approach of the Rough set theory that can be used for framing new decision rules. The application of this approach may be used extensively in the fields of knowledge discovery, data mining or any other field concerning attribute reduction and feature selection. As a direction for future

research attempts may be made towards testing this method using some large databases and comparing this method with some others existing methods.

REFERENCES

[1] Pal S.K., Skowron (Eds.), A. 1999. Rough Fuzzy Hybridization: A new trend in decision making. Springer-Verlag, Berlin.

[2] Pawlak, Z. 1982. Rough sets, International Journal of Computer and Information Sciences 11: 341–356.

[3] Pawlak, Z. 1991. Rough Sets: Theoretical Aspects of Reasoning about Data, System Theory, Knowledge Engineering and Problem Solving, vol. 9, Kluwer Academic Publishers, Dordrecht, The Netherlands.

[4] Han, Jiawei., Kamber, Micheline. 2001. Data Mining: Concepts and Techniques. San Franciso CA, USA, Morgan Kaufmann Publishers.

[5] Ramakrishnan., Naren and Grama, Ananth Y. 1999. Data Mining: From Serendipity to Science. IEEE Computer, August 1999: 34-37.

[6] Williams, Graham J. and Simoff, Simeon J. 2006. Data Mining Theory, Methodology, Techniques, and Applications (Lecture Notes in Computer Science/ Lecture Notes in Artificial Intelligence), Springer.

[7] Hand, D.J., Mannila, H., & Smyth, P. 2001. Principles of Data Mining. Cambridge, MA: MIT Press.

[8] Hand, D.J., Blunt, G., Kelly, M.G. & Adams, N.M. 2000. Data mining for fun and profit. Statistical Science, 15, 111-131.

[9] Glymour, C., Madigan, D., Pregibon, D., and Smyth, P. 1996. Statistical inference and data mining. Communications of the ACM, 39(11):35-41.

[10] Hastie, T., Tibshirani, R., & Friedman, J.H. 2001. Elements of statistical learning: data mining, inference and prediction. New York: Springer Verlag.

