International Journal of Mathematical Archive-2(9), 2011, Page: 1545-1549
Available online through www.ijma.info ISSN 2229 – 5046

A NOVEL APPROACH TO ATTRIBUTE REDUCTION UNDER DATA MINING PROCESS USING ROUGH SET THEORY

[1]Madhu.G*

Department of Information Technology, VNR Vignana Jyothi Inst of Engg & Technology, Batchupally, Nizampet (S. O.), Hyderabad-90, INDIA
E-mail: madhu_g@vnrvjiet.in
&
[2]E. Keshava Reddy

Department of Mathematics, Jawaharlal Nehru Technological University, Anantapur-515 002, A.P., INDIA
E-mail: keshava_e@rediffmail.com

(Received on: 01-05-11; Accepted on: 09-09-11)

ABSTRACT

Data mining is the process of selecting, exploring and modeling large amounts of data to uncover the previously unknown patterns. Data mining has gradually become an important and active area of research because of theoretical challenges and real-world applications associated with the problem of extracting interesting unknown patterns from large repositories. Attribute reduction has become an important pre-processing task to reduce the complexity of the data mining task. In rough set theory, accuracy and roughness are used to characterize uncertainty of a set and approximation accuracy is employed to depict accuracy of a rough classification. In this paper, we describe a novel approach for attribute reduction to Data mining process using rough set theory. Which is based on attribute reducts on the granular view of the information system.

Key words: *Attribute, Data mining, Information systems, Rough sets.*

---

1 INTRODUCTION

Data mining is the process of selecting, exploring and modeling large amounts of data to uncover previously unknown patterns [1]. Knowledge discovery aims to extract high-level knowledge or create a high-level description from real-world data sets [2]. Soft computing techniques, involving neural networks, genetic algorithms, fuzzy sets, and rough sets are mostly widely used in the data mining phase of the overall Knowledge Discovery (KD) process. Recently, rough set theory developed by Pawlak in [3] has become a popular mathematical framework in areas such as pattern recognition, image processing, feature selection, neural computing, conflict analysis, decision support, data mining and knowledge discovery process from large data sets [4] [5] [6]. Advancing statistical methods and machine learning techniques have played important roles in analyzing high dimensional data sets for discovering patterns hidden in it. But the ultra high dimensionalities of such datasets make the mining still a nontrivial task. Hence attribute reduction/ dimensionality reduction is an essential data preprocessing task for such data analysis, to remove the noisy, irrelevant or misleading features to produce a minimal feature subset. Attribute reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality that corresponds to the intrinsic dimensionality of the data [7]. Attribute reduction [8] is a key problem in rough set theory based on knowledge acquisition, and many researchers proposed some algorithms for attribute reduction [9][10]. Rough Sets theory is very broad application area, such as expert systems, decision support systems, machine learning, pattern recognition, data mining, artificial intelligence and so on [11]. Rough Sets theory applies to data mining supplying the mathematics tool for dealing with uncertain knowledge [12] [13]. Liu Qing and Zeng Huanglin studied the characteristic and application of Rough Sets theory [14] [15]. Yang Shanlin discussed the process of data analyzing based on data mining of Rough Sets, and proposed the application of this method to decision support system [16].

2 PRELIMINARIES

In this section, we define some basic rough set notations will be introduced, and several attribute reduct definitions will be described [17]. First we recall the concept of information systems. An information system is a pair $S = (U, A)$,

---

**Corresponding author: Madhu.G*, *E-mail: madhu_g@vnrvjiet.in*

where U is a non-empty finite set of objects, A is a non-empty finite set of attributes. U / A is all the equivalence classes of A.

2.1.1 Def.1 Indiscernible relation:

In a data set $A = \{A_1, A_2, ...A_n\}$ of attribute and consisting of objects, the indiscernibility relation $IND(X)$ for subset $X \subseteq A$ of attributes is

$$IND\ (X) = \left\{(o, o') \in t \times t \mid \forall A_i \in X, o[A_i] = o'[A_i]\right\} \quad (1)$$

Where $o[A_i]$ and $o'[A_i]$ are attributes values of objects o and o', respectively. That means objects o and o', have the same value at attribute in A_i.

2.1.2 Def.2 Lower/Upper approximations:

The lower approximation $\underline{IND(Y, X)}$ and the upper approximation $\overline{IND(Y, X)}$ of IND(Y) by IND(X) are expressed by means of using indiscernible sets as follows:

$$\underline{IND\ (Y, X)} = \left\{o \in t \mid \exists o' S(X)_o \subseteq S(Y)_{o'}\right\} \quad (2)$$

$$\overline{IND(Y, X)} = \left\{o \in t \mid \exists o' S(X)_o \cap S(Y)_{o'} \neq \phi\right\} \quad (3)$$

When an object o takes imprecise values for some attributes, it does not always take the same actual value as another object o', even if both objects have the same expression. The degree in an indiscernibility degree of the object o with the object o'.

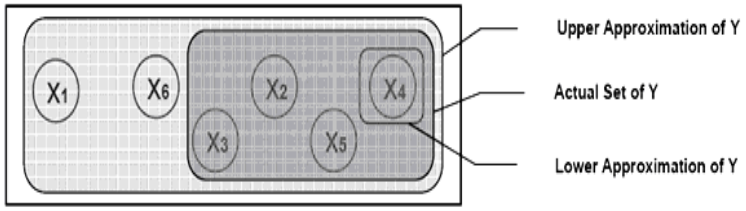

Fig: 1 Representation of lower and upper approximations

2.1.3 Information system

Based on Rough sets, an information system is a representation of data that describes some objects. Therefore the information system takes the form of relation table and the knowledge system with condition attribute is decision table. Suppose $S = (U, A, V, f)$ is knowledge system, where $S = (x_1, x_2, ...x_n)$ is a finite set of objects, $A = (a_1, a_2, ...a_n)$ is a finite set of attribute, here V is field composed of attribute A, $f : U \times A \rightarrow V$ is an information function, each element of U with a unique value that is *a* about V, $A = C \cup D$, where C is the condition attribute set, D is the decision attribute set. Example considers the descriptions of the several cars in Table 1.

Car	*Price*	*Size*	*Engine*	*Speed*
u_1	Max	Full	Max	Min
u_2	Min	Full	*	Min
u_3	*	Compact	*	Max
u_4	Max	Full	*	Max
u_5	*	Full	*	Max
u_6	Min	Full	Max	*

Table: 1 Information system about car

In table, where $U = \{u_1, u_2, u_3, u_4, u_5, u_6\}$ and $A = \{a_1, a_2, a_3, a_4\}$ with a_1 =Price, a_2 =Size, a_3 =Engine, a_4 = Speed. For any information system $S = (U, A)$ and an attribute subset $P \subseteq A$, we define a binary relation on U as follows:

$$SIM(P) = \{(u, v) \in U \times U \mid \forall a \in P, a(u) = a(v) ora(u) = * ora(v) = *\}. \quad (4)$$

Table.1 we obtain that $U / SIM(A) = \{S_A(u_1), S_A(u_2), S_A(u_3), S_A(u_4), S_A(u_5), S_A(u_6)\}$
Where $S_A(u_1) = \{u_1\}, S_A(u_2) = \{u_2, u_6\}, S_A(u_3) = \{u_3\}, S_A(u_4) = \{u_4, u_5\}, S_A(u_5) = \{u_4, u_5, u_6\}$, $S_A(u_6) = \{u_2, u_5, u_6\}$.

3. DATA MINING PROCESS USING ROUGH SETS

Data mining has attracted in the information industry and in society as a whole in recent years, due to the wide range amounts of data and the imminent need for turning such data into useful information and knowledge [18]. The Knowledge Discovery from Databases (KDD) is usually a multi-phase process involving numerous steps, like data preparation, preprocessing, search for hypothesis generation, pattern formation, knowledge evaluation, representation, refinement and management [19]. Recently Rough sets theory has become very popular in the research field of KDD. Rough sets theory has been considered as an intelligent mathematical tool for simplifying the KDD process [20].

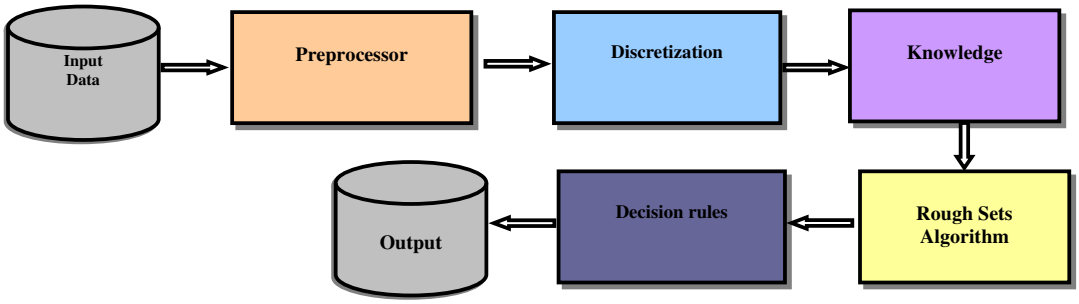

Fig: 2 Data mining process based on Rough sets.

The data mining process contains some stages, Input data, Preprocessor, Discretization, Knowledge, Rough sets algorithm, decision rules and output. The procedure is shown in the fig.2. All the steps have been connected with data evaluation. We now define a Reducation algorithm based on rough sets:
Algorithm:

Input: $S = (U, RED(C) \cup D, V, f), RED(C) = \{C_1, C_2, ...C_n\}$

Output: Decision table values S' *of* S *after reduction.*
1. $S' = (U, C \cup D, V' \leftarrow Null, f')$
2. for each attribute C_K *; do*
3. for each $x_i \in U$ *and* $C_k'(x_i) = Null$; *repeat*
4. if $\exists x_i((x_j \neq x_i) \wedge \forall C_l(C_l \neq C_k \wedge C_l(x_j) = C_l(x_i)) \wedge (D(x_j) \neq D(x_i))$
5. $C_k'(x_j) = C_k(x_j), C_k' = C_k(x_j)$
6. then
7. Output S'
8. end

4. EXPERIMENTS AND ANALYSIS

In this section, we describe our experiment results, in Information system $S = (U, A, V, f)$, where set $U = \{1,2,3,4,5,6,7\}$, $A = CUD$ $D = \{e\}$, $C = \{a, b, c, d\}$

U	*a*	*b*	*c*	*d*	*e*
1	1	1	0	0	1
2	1	0	0	0	0
3	1	0	1	0	1
4	1	1	0	1	0
5	2	0	1	2	2
6	1	1	2	0	2
7	0	2	2	2	2

Table: 2 Information system

Therefore after the reduction that $\{b,c,d\}$ is the attribute reduction of information system, expressed as $C^{'} = \{1,2,3\}$, if $N_1 = \{1100102\}$ and $N_2 = \{0002011\}$. Assume that $P_c \geq P_b \geq P_d$, that means $C = \{2,1,3\}$. From the table.2, the redundant attribute are eliminated and core attributes are preserved. Then the min decision rules are composed of core attributes without redundant attributes. Which is called reduced attribute table as shown in the below in table.3 & 4.

U/A	*a*	*b*	*c*	*d*	*e*
1,2	2	1	2	2	3
3	2	1	1	2	3
4	1	1	1	1	3
5	1	1	2	1	3
6,7	2	1	3	2	2
8	3	1	3	2	2
9,10	3	2	3	2	1
11,12,13	3	2	2	2	1

Table: 3 reduced attribute table.

U/A	*a*	*b*	*c*	*d*	*e*
1	2	1	2	2	3
2	2	1	1	2	3
3	1	1	1	1	3
4	1	1	2	1	3
5	2	1	3	2	2
6	3	1	3	2	2
7	3	2	3	2	1
8	3	2	2	2	1

Table: 4 Abstracted reduced attribute table.

U/A	*a*	*b*	*c*	*d*	*e*
1	2	-	2	2	3
2	2	-	1	2	3
3	1	-	-	1	3
4	1	-	-	1	3
5	-	-	3	2	2
6	-	1	-	2	2
7	-	2	-	2	1
8	-		-	2	1

Table: 5 reduced attribute table.

5. CONCLUSION

In this paper, we presented an approach for attribute reduction to Data mining process using rough set theory, this approach for the elimination of redundant data and the development of set of rules which can aid the attribute complexity. Also process the incomplete data is based on the lower and upper approximations and theory was defined as a pair of the two crisp sets to the approximations. We derived information table which can be generated the necessary decision rules for the aid to the reduced attribute tables. The integration of rough sets with other intelligent tools such as fuzzy sets and neural network for classification and rule generation in soft computing paradigm is the aim of our future work.

REFERENCE

[1] S. Mitra and T. Acharya, *Data Mining: Multimedia*, *Soft Computing*, *and Bioinformatics*, John Wiley & Sons, Inc., NY, USA, 2003.

[2] M. Kantardzic, *Data Mining: Concepts*, *Models*, *Methods and Algorithms*, John Wiley & Sons, Inc., New York, USA, 2002.

[3] Z. Pawlak, Rough sets, International Journal of Computer and Information Science.341–356, 1982.

[4] J. Bazan, J. F. Peters, A. Skowron, H. S. Nguyen, M. Szczuka, Rough set approach to pattern extraction from classifiers, Electronic Notes in Theoretical Computer Science 82 (4) (2003) 1–10.

[5] S. K. Pal, W. Pedrycz, A. Skowron, R. Swiniarski, Presenting the special issue on rough-neuro computing, Neurocomputing 36 (2001) 1–3.

[6] Z. Pawlak, A. Skowron, Rough sets and Boolean reasoning, Information Sciences 177 (2007) 41–73.

[7] Maaten L. J. P., Postma E. O. and Herik H. J. van den 2007. "Dimensionality Reduction: A Comparative Review", *Tech. rep. University of Maastricht*.

[8] Wang G. Y. (ed): Rough Set Theory and knowledge Acquisition, Xi'an Jiaotong university press, Xi'an.2001.

[9] Bazan J. G. Dynamic reducts and statistical inference, sixth International Conference on Information Processing and Management of University in Knowledge based systems (IPMU'96), Vol.3. Page 1147-1152.

[10] Wroblewski J. Theoretical Foundations of Order-Based Genetic Algorithms, Fundamenta Informaticae, Vol. 28 (3-4), 1996.423-430.

[11] Vasant Dhar, Roger Stein. Intelligent Dicision Support Methods. *The Science of Knowledge Work*. Printice Hall, 1997.

[12] Wu Bing, Li Weixi Decision-Making Method Based on Rough Set Theory [J]. *Information Technique*, Vol.4, No.4, 2002, pp. 4-5.(Chinese).

[13] Li Wanqing, Ma Lihua,Meng Wenqing. Analysis of Procedure of Project Decision Based on D Mining of Rough Sets [J]. *WSEAS Transactions on Business and Economics*, Vol.3, No.10, 2006, pp. 661-666.

[14] Liu Qing. *Rough Sets and Rough Inference* [M]. Beijing: Science Press, 2001.

[15] Zeng Huanglin. *Rough Sets and Applications* [M].Chongqing University Press, 1996.

[16] Yang Shanlin. *Intelligent Decision Method and Intelligent Decision Support System* [M]. Science Press, 2005.

[17] M. Kryszkiewicz. Comparative Study of Alternative Types of Knowledge Reduction in Inconsistent Systems. International Journal of Intelligent Systems, 2001, Vol.16, p105-120.

[18] Jiawei Han, Micheline Kamber "Data mining: concepts and techniques", Elsevier, second edition, 2007.

[19] Hrudaya Ku. Tripathy, B. K. Tripathy, and Pradip K. Das "An Intelligent Approach of Rough Set in Knowledge Discovery Databases", World Academy of Science, Engineering and Technology 35 2007.

[20] Madhu. G, G.Suresh Reddy, Dr. C. Kiranmai "Hypothetical Description for Intelligent Data Mining", International Journal on computer Science and Engineering, Volume 2 Issue 7, 2010.page., 2349-2352,

****

