REDUCT GENERATION FOR THE INCREMENTAL DATA USING ROUGH SET THEORY

Shampa sengupta[1], Asit Kumar Das[2]

[1]Department of Information Technology, MCKV Institute of Engineering,
Liluah, Howrah – 711 204, West Bengal, India
shampa2512@yahoo.co.in
[2]Department of Computer Science and Technology, Indian Institute of
Engineering, Science and Technology, Shibpur, Howrah – 711 103,
West Bengal, India
akdas@cs.becs.ac.in

ABSTRACT

In today's changing world huge amount of data is generated and transferred frequently. Although the data is sometimes static but most commonly it is dynamic and transactional. New data that is being generated is getting constantly added to the old/existing data. To discover the knowledge from this incremental data, one approach is to run the algorithm repeatedly for the modified data sets which is time consuming. The paper proposes a dimension reduction algorithm that can be applied in dynamic environment for generation of reduced attribute set as dynamic reduct. The method analyzes the new dataset, when it becomes available, and modifies the reduct accordingly to fit the entire dataset. The concepts of discernibility relation, attribute dependency and attribute significance of Rough Set Theory are integrated for the generation of dynamic reduct set, which not only reduces the complexity but also helps to achieve higher accuracy of the decision system. The proposed method has been applied on few benchmark dataset collected from the UCI repository and a dynamic reduct is computed. Experimental result shows the efficiency of the proposed method.

KEYWORDS

Dimension Reduction, Incremental Data, Dynamic Reduct, Rough Set Theory.

1. INTRODUCTION

In today's e-governance age, everything is being done through electronic media. So huge data is generated and collected from various areas for which proper data management is necessary. Retrieval of some interesting information from stored data as well as time variant data is also a very challenging task. Extraction of meaningful and useful data pattern from these large data is the main objective of data mining technique [1]. Data mining techniques basically uses the concept of database technology [2] and pattern recognition [3] principles. Feature selection [4] and reduct generation [5] are frequently used as a pre-processing step to data mining and knowledge discovery [6, 7]. For static data, it selects an optimal subset of features from the feature space according to a certain evaluation criterion. In recent years, dimension of datasets are growing rapidly in many applications which bring great difficulty to data mining and pattern recognition. As datasets changes with time, it is very time consuming or even infeasible to run

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 291–299, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4529

repeatedly a knowledge acquisition algorithm. Rough Set Theory (RST) [8, 9, and 10], a new mathematical approach to imperfect knowledge, helps to find the static as well as dynamic reduct. Dynamic reducts can put up better performance in very large datasets as well as enhance effectively the ability to accommodate noise data. The problem of attribute reduction for incremental data falls under the class of Online Algorithms and hence demands a dynamic solution to reduce re-computation. Liu [11] developed an algorithm for finding the smallest attribute set of dynamic reducts with increase data. Wang and Wang [12] proposed a distributed algorithm of attribute reduction based on discernibility matrix and function. Zheng et al. [13] presented an incremental algorithm based on positive region for generation of dynamic reduct. Deng [14] presented a method of attribute reduction by voting in a series of decision subsystems for generation of dynamic reduct. Jan G. Bazan et al. [15] presented the concept of dynamic reducts to solve the problem of large amount of data or incremental data.

In the proposed method, a novel heuristic approach is proposed to find out a dynamic reduct of the incremental dataset using the concept of Rough Set Theory. To understand the concepts of dynamic data, a sample dataset is divided into two sub sets considering one as old dataset and other as new dataset. Using the concept of discernibility matrix and attribute dependency of Rough Set Theory reduct is computed from old dataset. Then to handle the new data or incremental data, previously computed reduct is modified wherever changes are necessary and generates dynamic reduct for the entire system. The details of the algorithm are provided in subsequent section.

The rest of the paper is organized as follows: Basic Concepts of Rough Set Theory is described in section 2. Section 3 demonstrated the process of generation of dynamic reduct and Section 4 shows the experimental result of the proposed method. Finally conclusion of the paper is stated in section 5.

2. Basic Concepts of Rough Set Theory

The rough set theory is based on indiscernibility relations and approximations. Indiscernibility relation is usually assumed to be equivalence relation, interpreted so that two objects are equivalent if they are not distinguishable by their properties. Given a decision system DS = (U, A, C, D), where U is the universe of discourse and A is the total number of attributes, the system consists of two types of attributes namely conditional attributes (C) and decision attributes (D) so that $A = C \cup D$. Let the universe $U = \{x_1, x_2... x_n\}$, then with any $P \subseteq A$, there is an associated P-indiscernibility relation IND(P) defined by equation (1).

$$IND(P) = \{(x, y) \in U^2 | \forall a \in P, a(x) = a(y)\} \qquad (1)$$

If $(x, y) \in$ IND (P), then x and y are indiscernible with respect to attribute set P. These indistinguishable sets of objects, therefore define an indiscernibilty relation referred to as the P-indiscernibility relation and the class of objects are denoted by $[x]_P$.

The lower approximation of a target set X with respect to P is the set of all objects which certainly belongs to X, as defined by equation (2).

$$\underline{PX} = \{x | [x]_P \subseteq X\} \qquad (2)$$

The upper approximation of the target set X with respect to P is the set of all objects which can possibly belong to X, as defined by equation (3)

$$\overline{PX} = \{x | [x]_P \cap X \neq \emptyset\} \qquad (3)$$

As rough set theory models dissimilarities of objects based on the notions of discernibility, a discernibility matrix is constructed to represent the family of discernibility relations. Each cell in a discernibility matrix consists of all the attributes on which the two objects have the different values. Two objects are discernible with respect to a set of attributes if the set is a subset of the corresponding cell of the discernibility matrix.

(a) Discernibility Matrix and Core

Given a decision system DS = (U, A, C, D), where U is the universe of discourse and A is the total number of attributes. The system consists of two types of attributes namely conditional attributes (C) and decision attributes (D) so that A = C ∪ D. Let the universe U = {x_1, x_2... x_n}, then discernibility matrix M = (m_{ij}) is a |U| × |U| matrix, in which the element m_{ij} for an object pair (x_i, x_j) is defined by (4).

$$m_{ij} = \{a \in C : a(x_i) \neq a(x_j) \wedge (d \in D, d(x_i) \neq d(x_j))\} \qquad (4)$$

where, i, j = 1, 2, 3... n

Thus, each entry (i, j) in the matrix S contains the attributes which distinguish the objects i and j. So, if an entry contains a single attribute say, A_s, it implies that the attribute is self sufficient to distinguish two objects and thus it is considered as the most important attribute, or core attribute. But in reality, several entries may contain single attribute, union of which is known as core CR of the dataset, as defined in (5).

$$CR = \cup\{m_{ij} | m_{ij} \neq \emptyset \text{ and } |m_{ij}| = 1, \forall\, i, j = 1, 2, \ldots, n\} \qquad (5)$$

(b) Attribute Dependency and Reduct

One of the most important aspects of database analysis or data acquisition is the discovery of attribute dependencies; that establishes a relationship by finding which variables are strongly related to which other variables. In rough set theory, the notion of dependency is defined very simply. Assume two (disjoint) sets of attributes, P and Q, and inquire what degree of dependency is present between them. Each attribute set induces an (indiscernibility) equivalence class structure. Say, the equivalence classes induced by P is $[x]_P$, and the equivalence classes induced by Q is $[x]_Q$. Then, the dependency of attribute set Q on attribute set P is denoted by $\gamma_P(Q)$ and is given by equation (6).

$$\gamma_P(Q) = \frac{\sum_{i=1}^{N} \underline{|PX_i|}}{|U|} \qquad (6)$$

Where, Q_i is a class of objects in $[x]_Q$; ∀ i = 1, 2, …, N.

A reduct can be thought of as a sufficient set of attributes to represent the category structure and the decision system. Projected on just these attributes, the decision system possesses the same equivalence class structure as that expressed by the full attribute set. Taking the partition induced by decision attribute D as the target class and R as the minimal attribute set, R is called the reduct if it satisfies (7). In other words, R is a reduct if the dependency of decision attribute D on R is exactly equal to that of D on whole conditional attribute set C.

$$\gamma_R(D) = \gamma_C(D) \qquad (7)$$

The reduct of an information system is not unique. There may be many subsets of attributes which preserve the equivalence-class structure (i.e., the knowledge) expressed in the decision system.

(c) Attribute Significance: Significance of an attribute a in a decision table *A= (U, CUD)* (with the decision set *D*) can be evaluated by measuring the effect of removing of an attribute a$\in$C from the attribute set C on the positive region. The number γ(C, D) expresses the degree of dependency between attributes C and D. If attribute 'a' is removed from the attribute set C then the value of (γ(C, D)) will be changed.

So the significance of an attribute a is defined as

$$\sigma^{a}_{(C,D)} = \frac{\gamma(C,D) - \gamma(C - \{a\}, D)}{\gamma(C,D)} \qquad (8)$$

(d) Dynamic Reduct: The purpose of dynamic reducts is to get the stable reducts from decision subsystems. Dynamic reduct can be defined in the following direction.

Definition 1: If *DS* = (*U*, *A*, *d*) is a decision system, then any system *DT* = (*U'*, *A*, *d*) such that *U'* ⊆ *U* is called a subsystem of *DS*. By P (*DS*) we denote the set of all subsystems of *DS*. Let *DS* = (U, *A*, *d*) be a decision system and *F* ⊆ P (*DS*). By *DR* (*DS*, *F*) we denote the set *RED (DS)* $\cap \bigcap_{DT \in F}$ *RED (DT).* Any elements of *DR* (*DS*, *F*) are called an F-dynamic reduct of *DS*.
So from the definition of dynamic reducts it follows that a relative reduct of DS is dynamic if it is also a reduct of all sub tables from a given family of F.

Definition 2: Let *DS* = (U, *A*, *d*) be a decision system and *F* ⊆ P (*DS*). By *GDR* (*DS*, *F*) we denote the set

$$\bigcap_{DT \in F} RED\,(DT)$$

Any elements of *GDR* (*DS*, *F*) are called an F generalized dynamic reduct of *DS*. From the above definitions of generalized dynamic reduct it follows that any subset of *A* is a generalized dynamic reduct if it is also a reduct of all sub tables from a given family F.

Time complexity of computation of all reducts is NP-Complete. Also, the intersection of all reducts of subsystems may be empty. This idea can be sometimes too much restrictive, so more general notion of dynamic reducts are described. They are called (F, ε) dynamic reducts, where ε > 0. The set *DR (DS, F)* of all (F, ε) dynamic reducts is defined by

$$DR_{\varepsilon}^{(DS)} = \{C \in RED\,(DS, d): \frac{card(DT \in F : C \in red(DT, d))}{card\,(F)} \geq 1 - \varepsilon\}$$

3. Dynamic Reduct Generation Using Rough Set Theory

Various concepts of rough set theory like discernibility matrix, attribute significance and attribute dependency are applied together to compute dynamic reducts of a decision system. The term dynamic reduct is used in the sense that the method computes a set of reducts for the incremental data very quickly without unnecessarily increasing the complexity since they are sufficient to represent the system and subsystems of it. Based on the discernibility matrix M and the frequency value of the attributes, the attributes are divided [16] into the core set CR and noncore set NC for old subsystem DS_{old}. Next, highest ranked element of NC is added to the core CR in each iteration provided the dependency of the decision attribute D on the resultant set increases for the old subsystem ; otherwise it is ignored and next iteration with the remaining elements in NC is performed. The process terminates when the resultant set satisfies the condition of equation (7) for the old subsystem and is considered as an initial reduct RED_OLD. Then backward attribute removal process is applied for each noncore attribute x in the generated reduct RED_OLD, it is checked whether (7) is satisfied using RED_OLD – {x}, instead of R. Now if it is satisfied, then x is redundant and must be removed. Thus, all redundant attributes are removed and final reduct RED_OLD is obtained.

To generate the dynamic reduct, discernibility matrix is constructed for the new subsystem DS_{new} and frequency values of all conditional attributes are calculated. Now the previously computed reduct (RED_OLD) from the old dataset is applied to new dataset for checking whether it can preserve the positive region in the new data set i.e., whether the dependency value of the decision attribute on that reduct set is equal to that of the decision attribute on the whole conditional attribute set. If the condition is satisfied, then that reduct set is considered as dynamic reduct (DRED). Otherwise; according to the frequency values obtained using [16] of the conditional attributes, higher ranked attribute is added to the most important attribute set in each iteration provided attribute dependency of the resultant set increases and subsequently a reduct is formed after certain iteration when dependency of the decision attribute on the resultant set is equal to that of the decision attribute on the whole condition attribute set for the new subsystem. Then backward attribute removal process is applied for generation of final dynamic reduct of the system. In this process, significance value of each individual attribute is calculated using equation (8) except that most important attribute set in a reduct. If the significance value of a particular attribute is zero, then that attribute is deleted from the reduct. In this way, all redundant attributes are removed and finally dynamic reduct is generated by modifying the old reducts for the entire data.

The proposed method describes the attribute selection method for the computation of reducts from old data and dynamic reduct set DRED for entire data considering incremental data.

Algorithm1 generates initial reduct for the old decision system DS_{old} = (U, A, C, D) and Algorithm2 generates dynamic reduct for the entire data, by considering the old data as well as incremental data.

Algorithm1: Initial_Reduct_Formation (DS_{old}, CR, NC)

Input: DS_{old}, the decision system with C conditional attributes and D decisions with objects x, CR, the core and NC, the non-core attributes

Output: RED_OLD, initial reduct

```
Begin
        RED_OLD = CR   /* core is considered as initial reduct*/
        NC_OLD = NC /* take a copy of initial elements of NC*/
       /*Repeat-until below forward selection to give one reduct*/
        Repeat
             x = highest ranked element of NC_OLD
             If (x = $\phi$) break   /*if no element found in NC*/
             If ($\gamma_{RED\_OLD \cup \{x\}}$ (D) > $\gamma_{RED\_OLD}$ (D))
                 {
                   RED_OLD = RED_OLD $\cup$ {x}
                   NC_OLD = NC_OLD - {x}
                 }
        Until ($\gamma_{RED\_OLD}$(D) = $\gamma_C$(D))
          // apply backward removal
        For each x in (RED_OLD – CR)
              If ($\gamma_{RED\_OLD - \{x\}}$(D) = = $\gamma_C$(D))
                    RED_OLD = RED_OLD - {x}
       Return (RED_OLD);
End
```

Algorithm2: Dynamic_Reduct_Formation (DS, C, D)

//An algorithm for computation of dynamic reducts for incremental data

Input: DS = {DS_{new}}, the new decision system with C conditional attributes and D decisions attribute and reduct RED_OLD obtained from 'Reduct Formation' algorithm for the old dataset (DS_{old}).

Output: Dynamic reduct (DRED), reduct of $DS_{old} \cup DS_{new}$

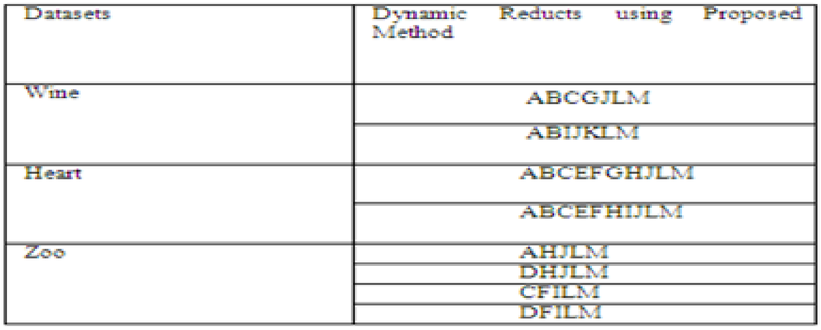
```
Begin
    If (($\gamma_{(RED\_OLD)}$ (D) = $\gamma_{(C)}$ (D))
        {
          DRED = RED_OLD
          Return DRED
        }

    Else {
        NC = C - RED_OLD
        CR = DRED  /*initial reduct is considered as core reduct of new system */
        Repeat
           DRED = RED_OLD   /* Old reduct is considered as core */
           x = highest frequency attribute of NC
```

```
 If ($\gamma_{DRED\cup\{x\}}(D) > \gamma_{DRED}(D)$)
            {
             DRED = DRED $\cup$ {x}
             NC = NC - {x}
            }

     Until ($\gamma_{DRED}(D) = \gamma_C(D)$)

     // apply backward removal

For each highest ranked attribute x in (DRED – CR) using (8)

      If ($\gamma_{DRED - \{x\}}(D) == \gamma_C(D)$)
          DRED = DRED - {x}
    Return (DRED);
  }  /* end of else*/
End
```

4. EXPERIMENTAL RESULTS

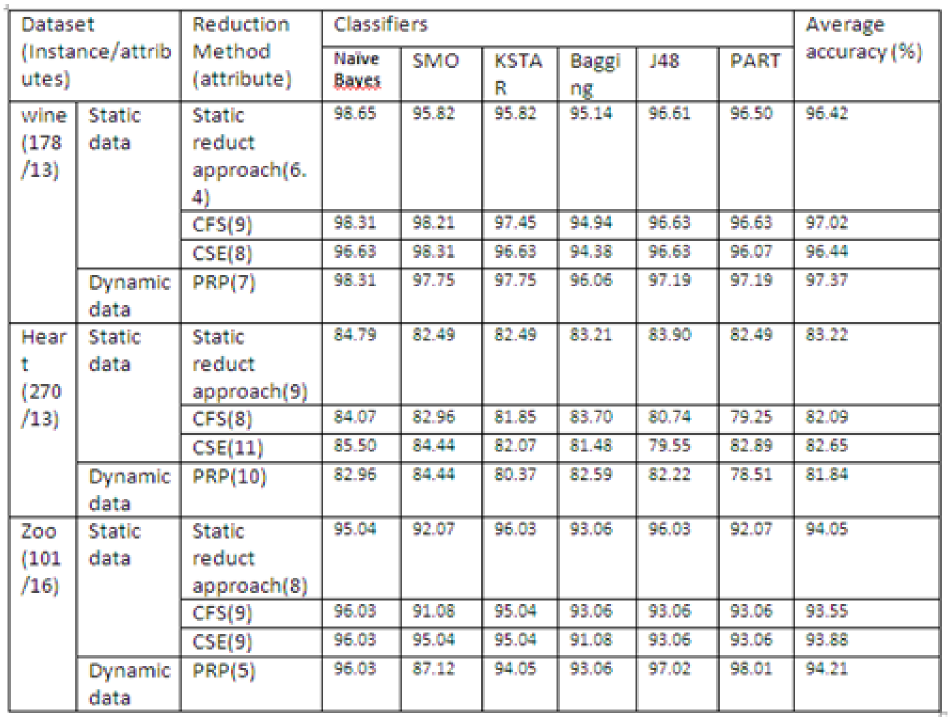
The method is applied on some benchmark datasets obtained from UCI repository ‘http: //www.ics.uci.edu/mlearn/MLRepository’. The wine dataset contains 178 instances and 13 conditional attributes. The attributes are abbreviated by letters A, B, and so on, starting from their column position in the dataset. In our method, for computation of dynamic reduct the wine dataset is divided into 2 sub tables considering randomly 80% of data as old data and other 20% of data is new data. Reduct is calculated for the old data using Algorithm1.Then based on previous reducts, the proposed algorithm worked on new data and generates two dynamic reducts {{ABCGJLM}, {ABIJKLM}} for the whole dataset. Similarly dynamic reducts are calculated for the heart and Zoo dataset. Reducts are also calculated for the modified data set using static data approach. All results are given in Table 1. Accuracies of the reduct of our proposed algorithm (PRP) are calculated and compared with existing attribute reduction techniques like ‘Correlation-based Feature Selection’ (CFS) and ‘Consistency-based Subset Evaluation’ (CSE), from the ‘weka’ tool [17] as shown in Table 2. The proposed method, on average, contains lesser number of attributes compared to CFS and CSE and at the same time achieves higher accuracy, which shows the effectiveness of the method.

Table 1. Dynamic reducts of datasets



Datasets	Dynamic Reducts using Proposed Method
Wine	ABCGJLM
	ABIJKLM
Heart	ABCEFGHJLM
	ABCEFHIJLM
Zoo	AHJLM
	DHJLM
	CFILM
	DFILM

Table 2. Classification accuracy of reducts obtained by proposed and existing method



Dataset (Instance/attributes)		Reduction Method (attribute)	Classifiers						Average accuracy (%)
			Naive Bayes	SMO	KSTAR	Bagging	J48	PART	
wine (178 /13)	Static data	Static reduct approach(6.4)	98.65	95.82	95.82	95.14	96.61	96.50	96.42
		CFS(9)	98.31	98.21	97.45	94.94	96.63	96.63	97.02
		CSE(8)	96.63	98.31	96.63	94.38	96.63	96.07	96.44
	Dynamic data	PRP(7)	98.31	97.75	97.75	96.06	97.19	97.19	97.37
Heart (270 /13)	Static data	Static reduct approach(9)	84.79	82.49	82.49	83.21	83.90	82.49	83.22
		CFS(8)	84.07	82.96	81.85	83.70	80.74	79.25	82.09
		CSE(11)	85.50	84.44	82.07	81.48	79.55	82.89	82.65
	Dynamic data	PRP(10)	82.96	84.44	80.37	82.59	82.22	78.51	81.84
Zoo (101 /16)	Static data	Static reduct approach(8)	95.04	92.07	96.03	93.06	96.03	92.07	94.05
		CFS(9)	96.03	91.08	95.04	93.06	93.06	93.06	93.55
		CSE(9)	96.03	95.04	95.04	91.08	93.06	93.06	93.88
	Dynamic data	PRP(5)	96.03	87.12	94.05	93.06	97.02	98.01	94.21

5. CONCLUSION

The paper describes a new method of attribute reduction for incremental data by using the concepts of Rough Set theory. Even if the data is not completely available at a time, i.e it keeps arriving or increasing, the algorithm can find the reduct of such data without recomputing the data that has already arrived. The proposed dimension reduction method used only the concepts of rough set theory which does not require any additional information except the decision system itself. Since, reduct generation is a NP-complete problem, so different researchers' use different heuristics to compute reducts used for developing classifiers. Dynamic reducts are very important for construction of a strong classifier. A future enhancement to this work is to formation of classifiers from dynamic reduct sets and finally ensemble them to generate an efficient classifier.

REFERENCES

[1] Han, and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2001.

[2] Handbook of Research on Innovations in Database Technologies and Applications: Current and Future Trends Viviana E. Ferraggine , Jorge H. Doorn , Laura C. Rivero, ISBN-10: 1605662429 ISBN-13: 978-1605662428

[3] Devijver, P.A., and Kittler, J. (1982) Pattern Recognition: A Statistical Approach Englewood Cliffs, NJ: Prentice Hall.

[4] Della Pietra,S., Della Pietra, V., and Lafferty, J. (1997) Inducing features of random fields. IEEE transactions on pattern Analysis and Machine Intelligence, 19(4),pp. 380-393.

[5] R. Jensen, QiangShen, "Fuzzy-Rough Attribute Reduction with Application to Web Categorization, Fuzzy Sets and Systems, Vol.141, No.3, pp.469-485, 2004

[6] N.Zhong and A. Skowron, “A Rough Set-Based Knowledge Discovery Process”, Int. Journal of Applied Mathematics and Computer Science. 11(3), 603-619, 2001. BIME Journal, Volume (05), Issue (1), 2005

[7] Ethem Alpaydin Introduction to Machine Learning.PHI, 2010

[8] Pawlak, Z.: “Rough sets.: International journal of information and computer sciences,” Vol, 11, pp. 341-356 (1982)

[9] Pawlak, Z.: “Rough set theory and its applications to data analysis,” Cybernetics and systems 29 (1998) 661-688, (1998)

[10] K. Thangavel, A. Pethalakshmi. Dimensionality reduction based on rough set theory : A review, Journal of Applied Soft Computing, Volume 9, Issue 1, pages 1 -12, 2009.

[11] Z.T,Liu.: “An incremental arithmetic for the smallest reduction of attributes” Acta Electro nicasinicia, vol.27, no.11, pp.96—98,1999

[12] J.Wang and J.Wang,”Reduction algorithms based on discernibility matrx:The order attributes method.Journal of computer Science and Technology,vol.16.No.6,2001,pp.489-504

[13] G.Y.Wang,Z.Zheng and Y.Zhang”RIDAS-A rough set based intelligent data analysis system” Proceedigs of the 1st International conference on machine Learning and Cybernatics,Beiing,Vol2,Feb,2002,pp.646-649.

[14] D.Deng,D.Yan and J.Wang,”parallel Reducts based on Attribute significance”, LNAI6401, 2010, pp.336-343.

[15] G..Bazan ,”Dynamic reducts and statistical Inference” Proceedigs of the 6th International conference on Information Processing and Management of uncertainity in knowledge based system,July 125, Granada,Spain,(2),1996pp.1147-1152

[16] Asit Kumar Das, Saikat Chakrabarty, Shampa Sengupta “Formation of a Compact Reduct Set Based on Discernibility Relation and Attribute Dependency of Rough Set Theory” Proceedings of the Sixth International Conference on Information Processing – 2012 August 10 - 12, 2012, Bangalore, Wireless Network and Computational Intelligence Springer pp 253-261.

[17] WEKA: Machine Learning Software, http://www.cs.waikato.ac.nz/~ml/