Comparative Review on Classical Rough Set Theory based Feature Selection Methods

R.K. Bania
Assistant Professor
Dept. Computer Application
North Eastern Hill University,
Meghalaya, Tura, India

.

ABSTRACT
Real world big data are uncertain and imprecise in nature. Receiving higher accuracy in data analysis from those data is an important and challenging task. For the analysis task significant and informative set of data is more required. Feature selection (FS) refers to a problem for selecting maximum relevant and less redundant data, which can provide a superior precision for data analysis purpose. Rough set theory (RST), one of the successful compelling components of soft computing is used for analyzing different data. It has been widely used as a mathematical tool for FS. This paper starts with an outline of the basic concepts of FS and RST.A theoretical comparative review on some existing RST methods are discussed with their pros and cons. Three RST based FS algorithms like Quickreduct (QR), Relative Reduct (RR) and Entropy based Reduct (EBR) are presented briefly. An experimental study on these three algorithms is carried out. Six public domain datasets available in UCI machine learning repository is analyzed for their performance.

Keywords
Feature selection, Rough set, Soft computing, Supervised, Filter, Dependency

1. INTRODUCTION
The problems which are related in the areas of machine learning [4], pattern recognition (PR) [4], data mining [2] maximum of the time has to deal with dataset of very big in size i.e. the dimensions in rows and columns of the dataset are large. Technically handling the big datasets of different categories and type [2] is not a computationally effortless task. There should be some mechanisms or techniques which can give remedy to this problem. Dimensionality reduction (DR) [3, 5] is one important technique in knowledge discovery from big datasets. The main task of DR is to reduce the size of the original data without affecting its internal structure. In pattern recognition area, objects or patterns are described by a set of dimensions called the features. The final task of PR is to assign specific labels to different objects. Depending on the structure of data, PR problems mainly leads to the development of unsupervised and supervised learning techniques [4]. In supervised learning PR problems, it allows inputting all features from the dataset but in reality some of the features may not be relevant to give a solution. FS is one of the main important processes of pattern recognition cycle for supervised learning which can deal with the problem of DR for big data. In common, to select the set of good relevant features which can provide a better result with a less possibility of information loss, in a problem domain gives the basic idea of FS [1, 3, 5].The main advantages of FS are that it reduces high computational effort and demanding superior precision. On the other hand, the important issues in real world big datasets are that the datasets can be imprecise, uncertain, and incomplete in nature [1, 7]. So data analysis becomes more complicated. Soft computing (SC) [7] is one term which combines many independent highly sound areas by which real world data problems can be tackled. Rough set theory (RST) [8, 10] is one of the strong component of SC for knowledge discovery from a dataset, where there is a high chance of imprecision, uncertainty. This theory was initially developed for a finite universe of discourse in which the knowledge base is a partition, which is obtained by any equivalence relation defined on the universe of discourse. By looking at some of the merits of FS and RST, authors are motivated to study RST based FS techniques for handling real world big datasets. In this work, the methods which are discussed are about the precise and imprecise labeled data handling problem with respect to FS technique i.e. about supervised FS using classical RST.

The remaining part of the paper is structured as follows: Section II describes the framework of FS. Section III explains the basic ideas of RST. Section IV describes some of the existing algorithms. Section V discusses the experimental results. Sections VI give Discussion and Section VII will conclude the paper.

2. THEORITICAL BACKGROUND
2.1 Dimensionality Reduction
Most big datasets in reality contain a certain amount of irrelevant and redundant data which will not give any contribution to data analysis. There is a high chance that they may mislead the overall process. DR is one step to find meaningful and useful features which can represent the overall dataset and remove irrelevant features. DR can be done by using proper techniques which fall in to a few categories: first one that transform or encode the basic intuitive meaning of the feature set known as feature extraction. In feature extraction, techniques like Principal component analysis (PCA) [2], independent component analysis (ICA), linear discriminant analysis (LDA) [2] are used to find a linear transformation DR. These techniques are used for linearly correlated data. These methods use the Euclidean structures of data points which are calculated. A main drawback of these methods is that the constructed features do not maintain the true meaning and also it requires complex computations. Second one that tries to preserve semantics of the original data i.e. the FS [14, 15].

2.2 Feature Selection
The principal part of FS is to determine a minimal good feature subset from a problem domain and attempt to maximize the performance of a given function of predictive accuracy in representing the original features under consideration. Let us consider a finite feature set $X=(x_1, x_2\ldots$

x_S) of S-dimension of features with a given output class label $l_c \epsilon L$, where $L= (l_1,l_2 \ldots l_C)$ is the set of the output class labels. The minimal subset representation of X say, $M=(x_1,x_2,..,x_R)$where $R \leq S$which maintain higher accuracy level of the same class label l_c as the original set X optimizing a criterion function f(M). After getting the reduced set it is beneficial to use it in the real world mining problems which increase the efficiency and accuracy of resulting output. How much a feature or feature subset is useful can be studied by looking at its relevancy and redundancy [3,5]. A feature is said to be relevant if it is foretell of the decision features, otherwise it is irrelevant. A feature is considered to be redundant if it is highly correlated with other features. Hence, the search for a good feature subset involves finding those features that are highly correlated with the decision features, but are uncorrelated with each other.

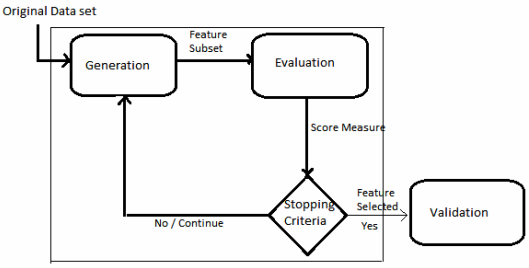

Fig. 1. Framework for feature selection

The general framework [1, 3], shown in the Fig.1, of FS process is integrated with feature subset generation and evaluation of the generated feature subset with a stopping criterion in an iterative way. Subset generation step requires an effective search method, in which the feature space is traversed in an attempt to locate valid feature subsets. In the literature three different kinds of search strategy are categorized, they are mainly: heuristic or informed, random and exponential or complete search methods [3]. The objective or evaluation function is to evaluate the goodness of the subsets. Evaluation function calculation will be either independent or dependent, which uses different criterion like, information, consistency, dependency, classification error measure etc. There is one stopping criteria which is checking in each iteration whether the FS process should continue its execution or not. There is another step some authors [1] integrate with the FS framework i.e. the validation process of the selected feature subset. It tries to test the validity of the selected subset with help of different tests, and comparing the results with previously established results. The features selection process can be categorized based on selection criteria into four main categories [1, 5]: Filter method independent in nature for subset selection from learning procedure; in Wrapper method feature subset selection is dependent on the learning procedure; in Embedded method feature selection is done during the learning process and lastly Hybrid method which takes the advantages of the other methods by combining them. Discussion here is based on the first category which uses feature evaluation criterion like dependency and consistency.

3. ROUGH SET BASED FEATURE SELECTION ALGORITHMS

3.1 Rough Set Theory

Rough set theory, introduced by Zdzislaw Pawlak in the early 1980s [8, 9, 14] can be regarded as a new mathematical tool to deal with imprecision and uncertainty. Since this mathematically sound concept proposed, it has been applied to many of the computational problems like hidden pattern recovery, data significance calculation, data reduction, rule extraction etc. RST concept is based on the assumption that some information is associated with every object in a dataset. The rough set is the approximation of a vague concept (set) by a pair of precise concepts, called lower and upper approximations. The lower approximation is a description of the domain objects which are known with certainty to belong to the subset of interest, whereas the upper approximation is a description of the objects which possibly belong to the subset.

Let I = (**U**, A) be an information system [8], where **U** is the universe with a non-empty set of finite objects, A is a non-empty finite set of condition attributes or features, and d is the decision attribute. Let $A \cap \{d\} = \emptyset$. The elements of the universe are called objects or instances. In the context RST, an equivalence relation is termed as an indiscernibility relation. There is a corresponding function for each $a \in A$, f_a: $\mathbf{U} \rightarrow V_a$ which is determined by the decision table and V_a is domain of a. For each non empty subset $R \subseteq A$, there is an associated indiscernible relation:

$$\mathrm{IND}(R) = \{(x,y) \in \mathbf{U}^2 \mid \forall a \in R, f_a(x) = f_a(y)\}$$

The partition of **U** generated by IND (R) is denoted as **U**/R. If (x, y) ∈ IND(R), then *x* and *y* are indiscernible by attributes from R. The equivalence classes of the R-indiscernability relation are denoted as $[x]_R$. Let $X \subseteq \mathbf{U}$ can be approximated using only the information contained within R by constructing the R-lower approximation $\underline{R}X$ and R-upper approximation RX of set X can be defined as:

$$\underline{R}X = \{ x \in \mathbf{U} \mid [x]_R \subseteq X\}$$

$$\overline{R}X = \{ x \in \mathbf{U} \mid [x]_R \cap X \neq \emptyset$$

It is such a tuple $(\underline{R}X, \overline{R}X)$ that is called a rough set. Let P, $Q \subseteq A$ be equivalence relations over **U**, then the positive, negative, and boundary regions can be defined as:

$$\mathrm{POSP}(Q) = \underline{R}X$$

$$\mathrm{NEGP}(Q) = \mathbf{U} - \overline{R}X$$

$$\mathrm{BNDP}(Q) = \overline{R}X - \underline{R}X$$

The positive region of the partition **U**/Q with respect to R, $\mathrm{POS}_R(Q)$, is the set of all objects of **U** that can be certainly classified to blocks of the partition **U**/Q by means of R. An important issue in data analysis is discovering Dependencies between attributes. A set of attributes depends totally on a set of attributes P, denoted by P⇒Q, if all attribute values from Q are identically determined by values of attributes from P. By using the positive region definition, dependency is defined in the following way:

$$z = \Upsilon_R(Q) = \frac{|\mathrm{POS}_R(Q)|}{|\mathbf{U}|}$$

When R is a set of condition attributes and Q is the decision, $\Upsilon_R(Q)$ is the quality of dependence. If z =1, Q depends totally on P; if $0 < z < 1$, Q depends partially on R, and if z= 0 then Q does not depend on R. The goal of attribute reduct is to remove redundant attributes so that the reduced set provides the same quality of precision as the original set. For this, the concept of a reduct is introduced and defined as a minimal subset Red of the initial attribute set A so that for

a given set of attributes D, $\gamma_{Red}(D) = \gamma_A(D)$. The reduct is the minimal subset if $\gamma_{Red-\{a\}}(D) \neq \gamma_{Red}(D)$ $\forall a \in Red$. This means that no attributes can be removed from the subset without affecting the dependency degree.

4. EXISTING RST BASED FS

In the literature there has been different approaches adopted for solving the DR problem with respect to FS. RST is well defined mathematical tool which gives the techniques for handling high dimensional data. Several methods are studied here with their advantages and disadvantages pointed out.

4.1 QuickReduct Algorithm

The Quickreduct(QR) algorithm[8] is a Dependency Function-based Reduction which added the forward elimination search to calculate a reduct set from the whole feature set. Procedure tries to find single reduct set without exhaustively generating all possible subsets. The lower approximation concept of the objects is calculated then dependency measure is calculated on each feature. The reduction of features is achieved by comparing equivalence relations generated by sets of attributes. As a whole the procedure starts off with an empty set and adds features in turn at a time, those that result in the greatest increase in the rough set dependency metric. It continues until the maximum possible value for the dataset is reached.

4.2 Relative Reduct Algorithm

Most existing rough set-based FS methods suffer from intensive computation of either discernibility functions or positive regions to find attribute reduct. This is another measure which is also used to discover rough set reducts. In Han et al. [11], a FS method which is based on an alternative dependency measure is presented. The authors replace the traditional rough set degree of dependency with an alternative measure. The degree of dependency is calculated as follows:

$$R = \left|\frac{U}{IND(Red)}\right| \qquad P = \left|\frac{U}{Red \cup D}\right|$$

$$RRed(D) = \frac{R}{P}$$

4.3 Entropy based Reduct Algorithm

Entropy based reduct Algorithm (EBR) [10] is based on the entropy heuristic employed by machine learning techniques such as C4.5. Method is mainly concerned with examining a dataset and determining those features that provide the most gain in information. The entropy of feature A (which can take values $a_1 \ldots a_n$) with respect to the conclusion C (of possible values $c_1 \ldots c_m$) is defined as:

$$H(C|A) = -\sum_{r=1}^{n} p(a_r) \sum_{s=1}^{m} P(c_s|a_r) \log_2 p((c_s|a_r)$$

Advantage of this algorithm is that it has user input threshold option. Algorithm search strategy for the best feature subset is stopped when the resulting subset entropy is equal to that of the entire feature set. For consistent data, the final entropy of the subset will be zero. It is interesting to note that any subset with entropy of 0 will also have a corresponding rough set dependency of 1. Hence, this technique can be used for finding rough set reduct if the data is consistent.

The main drawbacks of the above three algorithms are that they works fine to those dataset which are consistent and precise in nature. Below in Table 1, some other existing methods are shown. Column 4 and 5 describes their pros and cons of the methods respectively.

5. EXPERIMENTAL RESULTS

The datasets are taken from the UCI machine learning Repository. Table 2 shows the details of six datasets information. Column 2, 3 and 4 are object numbers, feature numbers and class numbers of the corresponding datasets. Three existing methods QR, RR and EBR has been implemented in C and in Linux environment which runs on a computer with the following Configuration: Processor: Intel® Core2TM Duo 2.00GHz, RAM: 4GB. The numbers of features selected and execution time (in seconds) by various algorithms are shown in Table 3 and 4 respectively.

It can be observed from Table 3 that QR and RR algorithm on the dataset index No [4] has less performance than the EBR. QR performance compared to EBR is almost similar, only with dataset index No [4] it has less performance. RR has also not better performance in selecting feature subset compared to EBR with dataset index No [4 and 6].From Table 4 also it can be observed that the execution time for RR is higher than QR and EBR for all datasets.

Table 1: Comparison between RST methods

Algorithm	***Evaluation Criterion***	***Search Method***	***Pros***	***Cons***
QR [8]	Dependency	Informed	Easy to understand and for medium complete dataset works fine.	Dependency measure between the attributes sometimes leads the search to a non minimal path. Not convergent to all real data. Poor termination.
Dynamic Reduct [9]	Dependency	Informed	Easy to understand and for medium size discrete dataset works fine.	Several subjective choice and use of thresholds. For large dataset computational cost high.
EBR [10]	Information	Informed	It works fine for consistent dataset. No threshold input.	Large scale and Inconsistence data cannot handle.
Relative Reduct [11]	Dependency	Informed	It gives an easy alternative dependency evaluation measure.	
Velayutham et. al.[16]	Dependency	Informed	Medium size complete data is handled.	Threshold requires and directly cannot works for continuous data. Incomplete data cannot handle.

Liang et. al.[12]	Dependency	Informed	Can deals with large scale data by dividing data into multi sub-table granularity view.	Computationally expensive and problem with missing data.
Jensen et. al.[14]	Information	Random	Small and medium size data is handled with optimal subset.	For large scale data Computation cost is high due to entropy measures.
Chen et.al.[13]	Information	Random	Medium size data works fine.	For large scale data, slow performance
Sugana et. al.[15]	Information	Random	No random parameters require. And works fine for medium size dataset.	Cannot deal with all types of continuous data and missing value dataset.

Table 1. Dataset Information

Index	Dataset	Object	Features	Class
1	Iris	150	4	3
2	Liver	345	6	2
3	Car	1728	6	4
4	Pima	768	8	2
5	Vote	435	16	2
6	Heart	270	13	2

Table 3: Comparison of feature selected by various methods

Index	**Dataset**	**Comparison of Feature Selected by three different Methods**		
		QR	***EBR***	***RR***
1	Iris	3[1,2,3]	3[1,2,3]	3[2,3,4]
2	Liver	3[1,2,5]	3[1,2,4]	3[3,4,5]
3	Car	6[1,2,3,4,5,6]	6[1,2,3,4,5,6]	6[1,2,3,4,5,6]
4	Pima	4[1,2,7,8]	3[1,2,7]	4[1,6,7,8]
5	Vote	12[1,2,3,4,6,8,9,11,13,14,15,16]	12[1,2,3,4,6,8,9,11,13,14,15,16	13[1,2,3,4,5,6,8,9,11,13,14,15,16]
6	Heart	4[1,4,5,6]	4[1,4,5,6]	5[5,6,11,12,13]

Table 4: Comparison of execution Time (in seconds)

Index	**Dataset**	**Comparison of Time(in seconds) by three different Methods**		
		QR	***EBR***	***RR***
1	Iris	0.7284	0.7451	0.7621
2	Liver	2.102	2.002	2.376
3	Car	4.103	3.993	4.601
4	Pima	10.231	10.877	13.097
5	Vote	1.903	1.802	2.102
6	Heart	3.320	3.119	4.091

6. DISCUSSIONS

It has been observed from the experimental result that on some dataset EBR and QR performance is preferably better than RR with respect to number of feature selection and execution time. On the other hand, in the literature various methods have been proposed for supervised FS using RST. In maximum of the works shown in Table 1, researchers has concentrate on dependency or information measure in the lower approximation part of the RST. To deals with inconsistent and imprecise data lower approximation based FS methods may not be directly suitable. Involving the information presents in the boundary region of RST and hybridization of RST can be one solution for these issues.

7. REFERENCES

[1] M. Dash and H. Liu, "Feature selection for classification". *Elsevier, Intelligent Data Analysis*, vol.3, pp.131-156, 1997.

[2] J. Han and M. Kanmber, "Data Mining concepts and techniques 2nd Edition", Morgan Kaufmann Publishers March, 2006.

[3] H. Liu and L.Yu, "Toward integrating feature selection algorithms for classification and clustering." *IEEE Transactions on Knowledge and Data Engineering*, vol.17, pp. 491-502, 2005.

[4] I. L. Kuncheva, "Combining pattern classifiers: methods and algorithms",*A Wiley-Interscience publication*, 2004.

[5] I. Guyon and A.Elisseeff, "An introduction to variable and feature selection", *Journal of Machine Learning. Res*, vol.3, pp.1157-1182, 2003.

[6] A.P. Englebrecht, "Computational intelligence an introduction 2nd Edition", *Wiley* 2007.

[7] R.Bello and L.J.Verdegay, "Rough sets in the soft computing environment", *Elsevier, Information Sciences*, vol.212, pp.1-14, 2012.

[8] A. Chouchoulas and Q. Shen, "Rough set-aided keyword reduction for text categorization". *Applied Artificial Intelligence*, vol.15, pp.843-873, 2001.

[9] J. Bazan, "Dynamic reducts as a tool for extracting laws from decision tables", In *Proceedings of the 8th Symposium on Methodologies forIntelligent Systems*, Lecture Notes in Artificial Intelligence vol.869, pp.346–355,1994.

[10] R. Jensen and Q. Shen, "Fuzzy-Rough attribute reduction with application to web categorization". *Fuzzy Sets and Systems*, vol.141, pp.469-485, 2004.

[11] J. Han, "Feature subset selection based on relative dependency between attributes", *Rough Sets and Current Trends in Computing: 4th International Conference*,

RSCTC, Uppsala, Sweden, pp.176–185, 2004.

[12] L. Liang, "An efficient rough feature selection algorithm with a multi-granulation view", *International Journal of Approximate Reasoning* vol.53, pp.912–926, 2012.

[13] Y Chen, "A rough set approach to feature selection based on ant colony optimization", *Pattern Recognition Letters* vol.31, pp.226–233, 2010.

[14] R. Jensen, "Performing feature selection with ACO", To appear in *Swarm Intelligence and Data Mining*, Springer SCI book series, 2006.

[15] N. Suguna and K. Thanushkodi, "A novel rough set reduct algorithm for medical domain based on bee Colony optimization", *Journal of Computing* vol.6, pp.49-54, 2010.

[16] C. Velayutham and K. Thangavel, "Improved Rough set algorithms for optimal attribute reduct" Journal of Electronic Science and Technology, Vol. 9 No.2, pp.108-117, 2011.