Survey on Feature Selection for Data Reduction

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

ABSTRACT
The storage capabilities and advanced in data collection has led to an information load and the size of databases increases in dimensions, not only in rows but also in columns. Data reduction (*DR*) plays a vital role as a data prepossessing techniques in the area of knowledge discovery from the huge collection of data. Feature selection (*FS*) is one of the well known data reduction techniques, which deals with the reduction of attributes from the original data without affecting the main information content. Based on the training data used for different applications of knowledge discovery, *FS* technique falls into supervised, unsupervised. In this paper an extensive survey on supervised *FS* technique describing the different searching approach, methods and application areas with an outline of a comparative study is covered.

Keywords
Data reduction, Feature selection, Filter, Wrapper, Embedded

1. INTRODUCTION
Knowledge becomes significant when it can be utilize efficiently in point of fact. Therefore knowledge management is progressively more being recognized as a key element in extracting its value. In the study of data mining [2], pattern recognition [5, 68], Machine learning [49] it is obvious to deal with the learning algorithms or procedures. The heart of the performance of learning algorithm is the prediction or the classification task. The task of the learning procedures is to induce a classifier by a finite numbers of training data that will be useful in classifying new instances. To induce or train the classifier an informative set of train data is required. On the other hand as in the various domains of real world computation, enormous technologies emerge the learning process has to deals with a very high dimensional real valued databases for data analysis. On those data there is high possibility of presents of noise, irrelevancy, and redundancy. Higher the dimensions processing and analysis become infeasible. So there should be some techniques to handle the high dimensional data so that computational complexity with respect to time and space of the learning and prediction algorithms reduces which can results a better learning accuracy. *DR* [2] is one of the data prepossessing techniques which are applied to handle the high dimensional data. The main concern of this technique is to reduce the size of the original data without affecting the original content. Reduction may be either in vertical columns of attributes wise or horizontal rows of instances wise by using some special techniques like *FS*, instance selection [15] for selecting the most meaningful information. *FS* is one of the best known data reduction techniques which can bring about a remedy to the problem of high dimensionality. The training data use for data mining applications and for learning approaches can be either labeled or unlabeled. It leads to the development of supervised, unsupervised and semi-supervised *FS* algorithms. Many researchers have studied these types of *FS* algorithms in separate ways. Supervised *FS* determines feature relevance by calculating feature's correlation or dependency [57] with the given class labels, whereas unsupervised *FS* exploits data variation and separability to evaluate feature relevance without the class labels. In this paper only supervised FS will be discussed.

2. FEATURE SELECTION
The *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 features set and second one that tries to preserve semantics. FS methods belong to the latter category, and first one is known as feature extraction [6]. In feature extraction, techniques like Principal component analysis (*PCA*) [2], independent component analysis (*ICA*), linear discriminant analysis (*LDA*) [2] are use to find a linear transformation *DR*. These techniques are used for linearly co related data. Euclidean structures of data points are calculated. Isomap, locally linear embedding (*LLE*) constructs are nonlinear *DR*. These methods can handle data of nonlinear relationship by calculating Euclidean structure. A main drawback of these methods is that the constructed features do not maintain the true meaning and also it requires complex computations. In *FS* smaller set of the original features is chosen based on a subset evaluation function. 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.

We assume that the *nxm* dimensional sample points are the input data for FS problem. The points are represented by a matrix R. In the classification task, we are given a pattern and the task is to classify it into one out of C classes. The number of classes, C, is assumed to be known a priori. Each pattern can be represented by a set of feature values, x(i), i = 1, 2, . . . m, which make up the m-dimensional feature vectors $x = [x(1), x(2), \ldots x(m)] \in R^m$ i.e. a feature vector is an m-dimensional vector of numerical features that represent a pattern. The classifier can be considering as a function f: $x \rightarrow C$. FS problem can be stated as, suppose $X = \{x(i) | i=1 \ldots m\}$ is an original feature vector with cardinality m. The objective of FS is to find $X' = \{x'(i) | i=1 \ldots n\}$ the new reduce feature vector, Where $n \leq m$ and to optimizing a criterion function $S(X')$.

The main aspect of *FS* is to find the correlation or dependency [57] i.e. the degree to which one feature subset is depends on another. That is why it is better to looks for high relevant and low redundant features [49]. It is required to remove the irrelevant and high redundant features. A feature is statistically relevant if its removal from the full set effects on accuracy. A feature is said to be redundant if one or more of

the other features are highly correlated with it. Hence, the search for a high-quality feature subset involves finding those features that are highly correlated with the class label features, but are uncorrelated with each other. In the literature lots of works can be found but to find informative and good features from the original features still it is a challenging task. For a problem domain of *S* dimension, numbers of features subset is 2^S-*1*. As *S* dimension increases to find out all the suitable possibilities become an exhaustive search. This leads to a non-polynomial time problem.

3. MOTIVATION

Since *FS* can bring a lot of advantages to the learning algorithms, firstly it reduces the computational complexity and it avoids over-fitting. Secondly, it provides robustness in the presence of noise, irrelevant and redundant feature and provides with elevated precision. By looking at all these advantages got motivated to study the *FS* technique for *DR*.

4. OBJECTIVES

Following are the objectives:

• To present a structured and comprehensive survey on existing methods.

• Analysis of existing methods in terms of their pros and cons.

• Highlights some real world applications, available tools and future directions related to FS process.

5. FRAMEWORK FOR FEATURE SELECTION

The general framework of *FS* process is integrated with feature generation and evaluation of the generated feature subset with a stopping criterion in an iterative way. Subset generation step require an effective search method, in which the feature space is traversed in an attempt to locate valid feature subsets. Then an objective or evaluation functions to evaluate the goodness of the subsets. There is one stopping criteria which is checking in each iteration to determine whether the *FS* process should continue its execution or not. Stopping criteria may be a specific constant value, or it may be the cardinality the subset features set, or a predefined number of iteration. Overall it is depends on the methods and task under consideration. There is another one 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 the help different tests, and comparing the results with previously established results. In the next sub sections different search methods and different approaches for the objective function are discussed.

5.1 Search Method

From the literature [1, 12, 21] it has been observed that the search strategy used in *FS* techniques poses some specific property and based on that search method are categorize into three groups. They are mainly: informed, random and complete or exponential. Under these roots several methods are classified. In this section a brief discussion is given regarding these searching methods.

5.1.1 Informed Search

It is one of the most commonly used search strategy for *FS* algorithms. In this method information about the problem i.e. the nature of the states, the cost of transformation from one state to another, the promise to taking a certain path and the characteristic of the goals are sometimes be used to help and guide the search more efficiently. It uses a heuristic function for the search problem and from a specific node to goal node it calculates the minimum cost. So, a search strategy which is better than another at identifying the most promising branches of a search-space is said to be more informed. The greedy hill climbing is an iterative process that starts processing with some arbitrary solution to a problem domain, and then tries to find a better solution by sequentially altering a single element from the solution. A heuristic function is used for finding the goal. It always looks for a better solution in a greedy manner, the process iterated until a better solution not come. Sequential forward selection, sequential backward elimination, combination of forward selection and backward elimination or bidirectional [48], sequential floating selection and *l*-plus and minus-*r* search [7], are some variations of this category. The property of sequential forward selection algorithms is that it starts processing with an empty set and during execution it adds features one by one. So as to maximize the intermediate criterion value until the required dimensionality is achieved. *SFS* (Sequential Forward Selection) results a minimal subset of *m* features: $X_m = add^m(\emptyset)$. *SBS*(Sequential Backward selection) which starts with a complete set of all the features and greedily remove features one by one.*SBS* results a minimal subset of *m* features: $X_m = remove^{(|M - m)}(M)$ Combination of both forward and backward elimination, leads to a selection of the best feature set and removes the irrelevant features from the remaining features. Among the three approaches widely used methods are *SFS* and *SBS*. But both *SFS* and *SBS* suffer from the nesting of feature subsets which leads to a less optimization ability. To overcome this problem either the Plus-*l*-Take away-*r* (also known as (*l, r*)) or generalized (*l, r*) algorithms [6] which involve successive augmentation and depletion process are employed. The extension of this idea leads to the basis of floating search approach. Sequential Forward Floating Selection (*SFFS*)[6], Sequential backward Floating Selection (*SBFS*), Oscillating Search(*OS*) [7] are some variations of this category.

The advantages of informed search methods are, they are quick in nature i.e. solution can be find within a limited period of time and often find a better solution, since more promising parts of the state-space can be examined, while ignoring the unpromising parts. Algorithms are easy to understand and easy for implementation. When we are looking for small number of features then forward selection is very much suitable. On the other hand backward elimination has the advantage that when it evaluates the relevance of a feature, it takes into concern all the other potential features. From demerits point of view, the greedy hill-climbing algorithms search uses minimum probable cost say $h(n)$ to the goal state as measure. This reduces the search time but the algorithm is neither complete nor optimal. It keeps only a single state in memory, but there is possibility to get stuck on local optima.

5.1.2 Randomized Search

Random search [46] has been used in many of the feature selection methods with different approaches to find an optimal feature subset. Random search algorithms always use some kind of randomness or probability in methods and the term metaheuristic is related with it. The search process starts with a randomly selected feature subset and proceeds in two different variations [1, 54].The instance-based methods [26] generate new subsets based on the current subset and uses heuristics to generate and update the subsets. Examples of instance-based algorithms includes simulated annealing (*SA*) [47], tabu search [47], genetic algorithm (*GA*) [9]. The other variation i.e. the model based methods basically rely on the sample distribution and the update parameters of the probability distribution. Ant colony optimization (*ACO*) [66],

stochastic gradient search are few examples of this category. *Advantages:* It provides a relatively good solution and statistically guarantees finding an optimal solution. Continuous and discrete global optimization problems are tackling in a way that is not possible for complete and informed search algorithms. *Disadvantages:* The trade off in some situation computational effort becomes high.

5.1.3 Complete or Exponential Search

This search strategies use no information about the likely path of the goal nodes i.e. the only information that it has is the initial state and no other information is known in priori. In a systematic manner it explores the nodes in some predetermined order or in a random way. From implementation point of view it is simple and will promise to gives a solution if it exists. But the cost is exponentially increased to the number of candidate features. If in the problem domain, *S* numbers of features available then the possible states it may have is 2^S. For a small number of problems size this method promise to provide good result. Some of the well known methods includes Branch and Bound [45] which is non exhaustive in nature, Depth first search and Breadth first search are uniformed.

5.2 Evaluation Criterion

In the earlier subsection as it has been mentioned that each newly generated feature subset should be evaluated but how feature subsets are evaluated is the single biggest differentiating factor among *FS* algorithms. For that there have an objective function for evaluating the subset of feature. It is based on independent or dependent criterion. How effective the feature subset is determined by using those criterions.

5.2.1 Independent Evaluation Criteria

In this criterion FS is done independently without any learning algorithm. In effect, only the relevant features are filtered in and irrelevant are filtered out before induction. Some of the popular independent criteria are distance measures, information measures, dependency measures, consistency measures.

Distance measures [30,65] are based on the statement that instances or objects of different classes are distant in the overall feature space. For the domain of two-class problem [2] e.g. feature *A* will be selected rather than feature *B*, if *A* induces a greater difference between the two-class conditional probabilities than *B*. *A* and *B* are said to be indistinguishable if the difference is zero. In distance measure, physical distances between objects are calculated using some function or metric which can differentiate between classes. Features that can support instances of the class to stay together are selected. The key concept is the assumption that instances of the same class must be closer than those in different class. *Information* measures [17,60] tries to measures the information or entropy gain from a feature *A*. The main motive is that from the given feature space *F*, which minimal feature subset *F* can gives extreme information gain. How much information gain is received form a feature *A* will be the difference between the prior uncertainty and expected next uncertainty using *A*. Feature *A* is selected than feature *B* if the information gain from *B* is smaller than that from *A*. *Dependency* measures [1,57] tries to measure how closely two features are associated or co-related with each other. They measure how feature *A* is dependent on the class label *C*. In most of the cases in *FS* for classification, how much the feature is dependent on the class label. If a feature *A* is highly dependent on another feature *B* than it is said to be as redundant feature i.e. feature *A* is preferred to another feature *B* if the association between feature *A* and class *C* is higher than the association between *B* and *C*. *Consistency* measures [13,36] attempt to find a minimal number of feature set that separate classes as consistently as the full set of features can. By calculating the pattern inconsistency rate, consistency measure is calculated. An inconsistent pattern can be defined as, if there are two instances such that they match feature values but their class labels are not match. e.g. (*a*, *b*, *c*1) and (*a*, *b*, *c*2), here two features take the same values for two instances but the class attribute varies. For a given feature set inconsistency rate is calculated and then with a user input threshold it is compared. If the value is less or equal than feature set is considered as consistent. In some filter method approaches [43] it is required to calculate the dependency measure of all the feature set with the class label if the measure is *1* then the training dataset is consistent else the training data is inconsistent.

5.2.2 Dependent Evaluation Criteria

A dependent criterion requires a predetermined classifier [2, 33]. The performance of the classifier is applied on the selected feature subset to determine which features will be selected. A classifier uses samples of instances for training and test set. More accurately the classifier will be trained by the training set implies a good results for the test set. The classification (prediction) accuracy refers to the capability of the classifier to correctly predict the class label of new or previously unseen data. Accuracy is the *percentage* of testing set correctly classified by the classifier. Now, the classifier error rate [12,33],is one of the dependent measures for calculating classifier accuracy. If *E* is the error rate for a feature subset, *T* is a threshold value, if $E < T$ than it will select the feature subset. This error rate holds a relationship with classification accuracy as the sum of predictive accuracy and error rate is *1*. Cross validation and bootstraps are most focusing and widely used error estimation techniques

6. FEATURE SELECTION METHODS

From convenient viewpoint, it is quite impossible to carry out a comprehensive study on all existing *FS* methods. Up to date, a large numbers of works has been published along with the research direction of *FS*. But with a handy approach, now turn to discuss about the methods with few review of some of the more famous *FS* algorithms. Based on the subset evaluation criteria there are basically three methods, they are namely: Filter, wrapper, embedded. Another method can also be found in some research works which combined the filter and wrapper methods together and it is known as the hybrid method.

6.1 Filter Method

This method used independent feature evaluation measure i.e. *FS* is done independently without involving any learning or induction algorithm. The popularity it has got in many domains as because of the independent nature of work. The advantages of using this method are that it is faster and tries to output optimal results. Filter approaches are generally employed where redundancy/irrelevance removal is the aim. The most common disadvantage of this method is that it ignores the interaction with the learning algorithm. In [30], the author proposed *RELIEF* algorithm, which is a one of the successful method using random sampling of records from the input datasets. It is distance based filter, for each data instance, the nearby example of the same class (nearest hit) and the nearby example from a different class (nearest miss) are selected. The extension version of this method can be found in [3]. The *FOCUS* [13] family of algorithms uses an exhaustive search in some situation when both the features

and class labels are binary. All feature subsets of increasing size are evaluated, until a sufficient set is establish. Another good method of filter category is *SIMBA*. It is a gradient-based optimization of the *NN* margin based feature selection criterion [63]. In [35], the author proposed a new pairwise constraint guided FS algorithm as constraint score and compare it with the well-known Fisher Score and Laplacian Score algorithms. *LVF* [14] is another filter method which used an alternative generation procedure by choosing the features randomly from the full feature set and accomplished the task by using Las Vegas algorithm [46]. Based on the entropy heuristics used in machine learning technique, another work was proposed which called as Entropy based reduction [56]. In [10], average Euclidean distance is used between instances in different classes as an evaluation function and genetic algorithms [15,16], as a search method. Using sequential search strategy for the task of association rule mining one novel method is proposed in [59]. A very recent worked based on information measure criteria, a unifying framework conditional likelihood maximization is proposed where instead of trying to define feature relevance indices, they derive it starting from a clearly specified objective function in [62]. Stochastic methods have also got more attention in the literature exploiting the merits of *ACO* ,*GA*, particle swarm optimization (*PSO*) [67] , ,differential evolution (*DE*). Two versions of *DE* based *FS* methods are presented in [70], where the desired feature subset size can be predefined by the user. There are extensively several works carried out with the help of rough set theory (*RST*) [53] and its extensions [56]. To overcome few drawbacks of *RST*, fuzzy set theory has been hybridized with it in [50, 51].

6.2 Wrapper Method

This method used dependent evaluation criteria i.e. *FS* is done with the involvement of any learning or induction algorithm. In the feature search space using an estimated accuracy from a learning or induction algorithm it searches for the suitability of the feature subset. The main advantage of this method is that it often got better results than filters. Reason is that it continuously maintains a precise communication between an induction algorithm and the training data. From the demerit point of view it is much slower than filter model because they must repeatedly re-run or call the induction algorithm. In every iteration typically it demands to evaluate a cross-validation. Finally there is also a higher risk of overfitting [14] than the filter methods. In the literature lots of works related wrapper methods are available. In [27], authors give an extensive review on the relation between optimal feature subset selection and relevance. They compare their propose wrapper method with *RELIEF* filter methods. 1-Nearest Neighbor's [29] extension instance-based algorithms [31] are merged with a beam search strategy by using a backward elimination can be found in [11]. *LVW* [34] which is a probabilistic approach that generates feature subsets in random fashion uses Las Vegas algorithm [46] by using a threshold named inconsistency rate. In [28], authors proposed a wrapper method for decision trees [2,32], with a search strategy, that is based on the strategy to add or remove features randomly, but also removing in each step all the features which were not included in the induced tree. Another wrapper Algorithm using Support Vector Machines (*SVM*) with kernel functions based on a sequential backward selection is proposed in [68].Using weights from the *SVM*, classification model is proposed in [69] and detail study for linear *SVM* is reviewed .For hyper spectral image classification purpose *GA* based method is proposed that uses *SVM* as a classifier in [55].

6.3 Embedded Method

The way *FS* and learning algorithm interact is different in an embedded method [37, 38]. In this method the learning part and the *FS* part cannot be divided. The internal structure of the class of functions and properties under consideration plays a crucial role. The search method is guided by the learning algorithm itself. The term embedded method is usually used to describe selection which is done automatically by the learning algorithm. The advantage of this method is that it has an interaction with the learning model and at the same time less computationally expensive than the other methods being so far. Overfiiting is also less prone to this method as it tends to have higher capacity of data involvement. In [43], the authors explore several recently developed *FS* technique for bioinformatics study with high dimensional data as an input. Overall review in that worked is a branch of embedded model. A small subset of gene selection from the broad patterns of gene expression data for cancer classification, there is very novel worked published in [39], using *SVM* as classifier. In [44], proposed an embedded method based on approximating the unknown distributions by a finite mixture of the densities of a product type using expectation maximization [2] method.

6.4 Hybrid Method

Algorithms under this method merged or combined the merits of more than one of the discussed methods in on solution to handle high dimensional data. The main characteristics of the algorithms are to focus on combining filter method to decide best subsets in efficient time and wrapper algorithms to achieve final best subset from the possible subsets. In the first stage using specific independent measure feature subset is selected and it is passed to second stage i.e. to wrapper method. Second stage process is repeated under particular conditions which includes different feature set sizes, dataset characteristics, classifiers and stopping criterion. From the merits point of view this method can involve the strength of the both models to improve the performance of *FS*. In [40, 41, 64], the authors combines the advantages of more than one of the *FS* methods to handle large dimension of data i.e. the hybrid model. In [42], authors proposed an interesting hybrid approach to combine the wrapper with the filter model called greedy randomized adaptive search procedure *GRASP*. In [52], an author has proposed an approach for neural network *FS* by combining filter and wrapper approach to solve the *DR* problem by reducing cross validation error. For classification problem in molecular biology in [61] author used a sequence of filters in three phases namely discritization, feature ranking and use of markov blanket filtering. Koller and Sahamis [31] filter and different three classifiers are also used there.

7. APPLICATIONS AND TOOLS

7.1 Applications

Some of the application areas for different FS methods are going to discuss in this section. As earlier also stated that in real world applications users often encounters many such problems related to high dimensional data where all of the features may not be the relevant for the task. *FS* is actively involved in some of the fields of research and development for last few years, namely in pattern recognition, machine learning, data mining and widely applied to many application area such as text categorization [16],remote sensing images [67], bioinformatics and gene expression [18,22],image retrieval [24],medical diagnosis [11], network intrusion detection [20] to name a few. Some of the illustrative applications are showcased below.

1.Text categorization: Text categorization is one kind of technique used to classify or categorize text information or

news stories, within few minutes. The purpose is to guide a search through the hyperlinks and hypertext so that user can discover attractive or interesting information on the web. So goal of this process is to categorize the documents into some classes. Several numbers of classification tasks on text categorization has been applied in [16]. The native space of documents consists of several numbers of features which sometimes prohibitively large for the learning algorithms. So, in this context *FS* is applied.

2. Image pattern classification: Hyperspectral sensors are mostly used to acquire remote sensing images. These images contain lots of spectral information and each pixel of the images is considered as pattern. To classify remote sensing images is not an easy task as it contains huge narrow and continuous bands of electromagnetic spectrum termed as attribute in each pixel. Huge numbers of bands are not always relevant for classification. Several works related to classification of remote sensing images with the help of *FS* has been carried out in [55, 67]. In the field of medical image research, mammography image classification [23] worked carried out and without removing some irrelevant feature predictive accuracy cannot be find out.

3. Bioinformatics and Genomic Analysis: The functional and structural data analysis of living creature's genome has increased in the recent years. These analyses have presented huge numbers of issues and challenges for data mining and pattern recognition area. Gene expression microarray analysis is a quickly growing technology which provides the chance to study the internal expression levels of thousands of genes in a an experiment. The size of the microarray data is very large for which computationally it become very expensive and rigorous. There are other several particular characteristics like noise and variability in the data leads to complications, so the *FS* process can give good result. In [24, 18, 61, 64] authors has applied *FS* in these area.

7.2 Software and Tools

In this section some of the popular software/tools names and links for *FS* technique is given. All these tools mentioned below are free for academic use.

1. WEKA http://www.cs.waikato.ac.nz/ml/weka
2. MLC++ http://www.sgi.com/tech/mlc
3. ROSETTA: //http://www.lcb.uu.se/tools/rosetta/
4. Spider http://www.kyb.tuebingen.mpg.de/bs/people/spider
5. GA-KNN http://dir.niehs.nih.gov/microarray/datamining
6. PCP http://pcp.sourceforge.net
7. GA-KNN http://dir.niehs.nih.gov/microarray/datamining/
8. Feature Selection Toolbox (*FST*) http://fst.utia.cz

8. DISCUSSIONS

Every family of *FS* methods has their own pros and cons. In general, the filter methods apply independent evaluation criterion like distance, information etc. without the involvement of any inducer or learning algorithm they are computationally become efficient. In most of the real world applications, frequently used *FS* algorithms are filters. The Wrapper methods involved a learning algorithm in spite of the independent criterion for subset evaluation process. Searching is done through the feature space using a learning algorithm. Estimated accuracy is calculated by the algorithm for each feature before it is added to or removed from the feature subset. It implies that learning algorithms are used to control the selection of feature subsets which are as a result better suited to the predetermined learning algorithm. Due to the necessity of the learning algorithm within the *FS* process, the wrapper methods are more computationally expensive than the filter methods. It requires to re-run when switching from one learning algorithm to another. Comparing the embedded model with the wrapper model, they are usually more efficient, since they look into the structure and used the properties o the involved learning model to guide feature evaluation and search. Hybrid *FS* algorithms can be defined easily to utilize the advantages of both filters and wrappers. In the process of search, in each algorithm step filter is used to reduce the number of candidates to be evaluated in wrapper. So, in ultimately it can be observed that filter method has some advantages over wrapper and embedded in some situations. And also the Hybrid model has performs well as the filter and wrapper is tightly coupled with it so, hybrid is also one of the winners.

9. CONCLUSION AND FUTURE DIRECTIONS

This survey provides a comprehensive study about the various approaches and methods related to FS technique. In this work instance selection is not included because it is not directly related to *FS.* It deals with the horizontal rows of the data. But they can be involved actively with *FS* technique. Researchers have proposed techniques regarding this issue. As various technology and data mining and retrieval techniques developed in various domain new problems are also arises related *FS.* Now a day's in maximum of the real world application like economics, medical research data are changing dynamically, i.e. group of instances and features are added in the data which may leads to previous information invalid or irrelevant. So in order to maintain effectiveness, it becomes necessary to establish good strategies for dynamic characteristic data. Another future direction is the unsupervised *FS* i.e. if decision class labels are not present in the data then how to perform *FS* process on those data for *DR*.

10. REFERENCES

[1] M. Dash and H. Liu. Feature Selection for Classification. Intelligent Data Analysis,vol.1,no.3,pp.131-156, 1997.

[2] J.Han and M.Kamber. Data Mining Concepts and Techniques 2nd Edition, Morgan Kaufmann Publishers March 2006.

[3] M. R. Sikonja and I.Kononenko. Theoretical and empirical analysis of Relief and ReliefF.Machine Learning, 53:23-69, 2003.

[4] L.Song, A.Smola, A.Gretton, K.Borgwardt, and J.Bedo. Supervised feature selection via dependence estimation. In International Conference on Machine Learning, 2007.

[5] I. L. Kuncheva. Combining pattern classifiers: methods and algorithms, Wiley-interscience Publication, 2004.

[6] P.Pudil, J.Novovicov, J,Kittler. Floating search methods in feature selection. Pattern Recognition. Letter,15(11), pp.1119-1125, 1994., 2005.

[7] P.Somol and P.Pudil. Oscillating search algorithms for feature selection. In ICPR 2000, Los Alamitos, CA, USA: IEEE Computer Society,volume 02, pp. 406-409,2000.

[8] J.Weston, A.Elisse, B.Schoelkopf, and M.Tipping. Use of the zero norm with linear models and kernel methods.Journal of Machine Learning Research, pp. 1439-1461, 2003.

[9] J. Bala, J. Huang, H. Vafaie, K.DeJong, and H.Wechsler. Hybrid learning using genetic algorithms and decision trees for pattern classification. In IJCAI (1), pp. 719-724, 1995.

[10] G.H.John, R.Kohavi, and K.Peger. Irrelevant feature and the subset selection problem. In W.W. Cohen and Hirsh H., editors,Machine Learning: Proceedings of the Eleventh International Conference,New Brunswick, N.J.Rutgers University. pp.121-129,1994.

[11] Z.Zhao and H.Liu. Semi-supervised feature selection via spectral analysis. In Proceedings of SIAM International Conference on Data Mining(SDM),2007.

[12] J.Doak. An Evaluation of Feature Selection Methods and Their Application to Computer Security. Technical report, University of California at Davis, Dept. of Computer Science, 1992.

[13] H.Almuallim and T.G.Dietterich. Learning Boolean Concepts in the Presence of Many Irrelevant Features, Artificial Intelligence, vol. 69,nos. 1-2,pp. 279-305,1994.

[14] H. Liu and R. Setiono, A Probabilistic Approach to Feature Selection-A Filter Solution, Proc. 13th Int'l Conf. Machine Learning, pp. 319-327, 1996.

[15] I.Guyon. A.Elisseeff. An introduction to variable and feature selection.J. Mach. Learn. Res., 3, pp.1157-1182.2003.

[16] E. Leopold and J. Kindermann. Text Categorization with Support Vector Machines. How to Represent Texts in Input Space? Machine Learning, vol. 46, pp. 423-444, 2002.

[17] D.A. Bell, H. Wang. A formalism for relevance and its application in feature subset selection, Machine Learning 41, pp.175-195. 2001.

[18] C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. In Proceedings of the Computational Systems Bioinformatics conference (CSB'03), pp. 523-529, 2003.

[19] C.Lai, M.J.T Reinders, L.J van't Veer, and L.F.A Wessels. A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. BMC Bioinformatics, 7:235, 2006.

[20] W. Lee, S. J. Stolfo, and K. W. Mok. Adaptive intrusion detection: A data mining approach. AI Review, vol. 14(6), pp.533-567, 2000.

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

[22] Y.Saeys, I.Inza, and P.Larraaga. A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19) pp.2507-2517, 2007.

[23] Y. Sun, C.F.Babbs, and E.J.Delp. A comparison of feature selection methods for the detection of breast cancers in mammograms: adaptive sequential floating search vs. genetic algorithm. Conf Proc IEEE Eng Med Biol Soc, 6:6532-6535, 2005.

[24] D. L. Swets and J. J. Weng. Efficient content-based image retrieval using automatic feature selection. In IEEE International Symposium On Computer Vision, pp. 85-90, 1995.

[25] L. Yu, C.Ding, and S.Loscalzo. Stable feature selection via dense feature groups. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008.

[26] D. W. Aha, D. Kibler, and M. K. Albert. Instance-based learning algorithms. Machine Learning, 6:37-66, 1991.

[27] R. Kohavi and G.H. John. Wrappers for Feature Subset Selection, Artificial Intelligence, vol. 97, nos. 1-2, pp. 273-324, 1997.

[28] R.Caruana and D.Freitag. Greedy attribute selection. In International Conference on Machine Learning, pp. 28-36, 1994.

[29] T.M. Cover and P.E. Hart. Nearest neighbor pattern classifier. IEEE Transactions on Information Theory, 13:21-27, 1967.

[30] K. Kira and L. Rendell. A practical approach to feature selection. In Proc. 9th International Workshop on Machine Learning, pp. 249-256, 1992.

[31] D.Koller and M.Sahami. Toward optimal feature selection. In International Conference on Machine Learning, pp. 284-292, 1996.

[32] J. R. Quinlan. Induction of decision trees. Journal of Machine Learning, vol-1 pp.81-106, 1986.

[33] J. R. Quinlan. C4.5: Programs for Machine Learning. The Morgan Kaufmann Series in Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA.1993.

[34] H.Liu and R. Setiono. Feature Selection and Classification-A Probabilistic Wrapper Approach, Proc. 9th Intl Conf. Industrial and Eng.Applications of AI and ES, T. Tanaka, S. Ohsuga, and M. Ali, eds., pp. 419-424, 1996.

[35] D. Sun, D. Zhang. Bagging Constraint Score for feature selection with pairwise constraints. Elsveir Pattern recognition pp.2106-2118, 2010.

[36] Q. Hu, H. Zhao, Z. Xie, and, D. Yu. Consistency based attribute reduction. PAKDD 2007, LNAI 4426, Yang (Ed.), vol.4426, pp.96-107, 2007.

[37] P.Pudil, P.Novovicov, N.Choakjarernwanit, J.Kittler. Feature selection based on approximation of class densities by finite mixtures of special type. Pattern Recognition, 28, pp.1389-1398. 1995.

[38] S. Mika, G. Ratsch, and K.-R. Muller. A Mathematical Programming Approach to the Kernel Fisher Algorithm. Advances in Neural Information Processing Systems,Cambridge, MA, USA, MIT Press. pp. 591-597, 2000.

[39] I.Guyon, J.Weston, S.Barnhill, V.Bapnik. Gene selection for cancer classification using support vector machines. Machine Learning, 46(1-3), pp. 389-422. 2002.

[40] M.Sebban, R.Nock. A hybrid filter/wrapper approach of feature selection using information theory. Pattern Recognition, 35, pp.835-846.2002.

[41] P.Somol,J.Novovicov,P.Pudil. Flexible-hybrid sequential floating search in statistical feature selection. In Structural, Syntactic, and Statistical Pattern

Recognition,Springer-Verlag, volume LNCS 4109 pp. 632-639,2006.

[42] M. A. Esseghir. Effective wrapper-filter hybridization through grasp schemata. In The 4th Workshop on Feature Selection in Data Mining, pp.45-54, 2010.

[43] S. Ma and J. Huang. Penalized feature selection and classification in bio informatics. Brief Bioinform, 9(5):392-403, Sep 2008.

[44] P.Pudil,P.Novovicov, N.Choakjarernwanit, J.Kittler, Feature selection based on approximation of class densities by finite mixtures of special type. Pattern Recognition, 28, pp.1389-1398. 1995.

[45] P.M .Narendra and K.Fukunaga, A Branch and Bound Algorithm for Feature Subset Selection, IEEE Trans. Computer, vol. 26, no. 9, pp. 917-922, Sept. 1977.

[46] A.A.Zhigljavsky. Theory of Global Random Search. Kluwer Academic.ISBN 0-7923-1122-1, 1991.

[47] J. Doak, An Evaluation of Feature Selection Methods and Their Application to Computer Security, technical report, University of California at Davis, Dept. Computer Science, 1992.

[48] H. Liu and H. Motoda. Feature Selection for Knowledge Discovery and Data Mining. Boston: Kluwer Academic, 1998.

[49] A.L. Blum and P. Langley. Selection of Relevant Features and Examples in Machine Learning, Artificial Intelligence, vol. 97, pp. 245-271,1997.

[50] R.Jensen and Q. Shen. Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough Based Approaches. IEEE Transactions on Knowledge and Data Engineering, 16(12): 1457-1471, 2004.

[51] R. Jensen and Q. Shen. Tolerance-based and Fuzzy-Rough Feature Selection, Proceedings of the 16th International Conference on Fuzzy Systems pp. 877-882, 2007.

[52] H. Yuan,S.S. Tseng,W. Gangshan and Z. Fuyan. A Two-phase Feature Selection Method using both Filter and Wrapper, IEEE transaction pp.132-136, 1999.

[53] Q. Shen, R. Jensen. Rough sets, their extension and applications International Journal of Automation and Computing Vol. No. 04(3), pp. 217-228, July 2007.

[54] S. Das. Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection, Proc. 18th Intl Conf. Machine Learning, pp. 74-81, 2001.

[55] L. Zhuo , J. Zheng , F. Wang, X. Li, B. Ai and J. Qian. A GA based wrapper feature selection method for classification of hyperspectral images using SVM, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B7. pp.397-402, 2008

[56] R. Jensen and Q. Shen. A Rough Set-Aided System for Sorting WWW Book- marks. In N. Zhong et al. (Eds.), Web Intelligence: Research and Development, pp. 95-105. 2001.

[57] M.A. Hall. Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning, Proc. 17th Int'l Conf. Machine Learning, pp. 359-366, 2000.

[58] L.Yu and H. Liu. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution, Proc. 20th International Conference. Machine Learning, pp. 856-863, 2003.

[59] S. Chawla. Feature selection, association rules network and theory building. In The 4th Workshop on Feature Selection in Data Mining, 2010.

[60] D.Koller and M. Sahami, Toward Optimal Feature Selection, Proc.13th Intl Conf. Machine Learning, pp. 284-292, 1996.

[61] E.P. Xing, M. Jordan, and R.M. Karp. Feature Selection for High-Dimensional Genomic Microarray Data,Proc. 15th International Conference. Machine Learning, pp. 601-608, 2001.

[62] G.Brown, A.Pocock, M.J.Zhao, M.Lujan, Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection, Journal of Machine Learning Research vol-13 pp.27-66.2012.

[63] R. Gilad-Bachrach, A. Navot, and N. Tishby. Margin based feature selection- theory and algorithms. In Proceedings of the International Conference on Machine LearningICML, pp. 43-50. ACM Press, 2004.

[64] Li Yeh Chuang, Chao Hsuan Ke, and Cheng Hong Yang. A Hybrid Both Filter and Wrapper Feature Selection Method for Microarray Classification, Proceedings of the International Multi Conference of Engineers and Computer Scientists 2008.

[65] J.Liang, R.Li, Y. Qian. Distance: A more comprehensible perspective for measures in rough set theory, Knowledge-Based Systems 27, pp.126-136,2012.

[66] M.Dorigo, Optimization, Learning and Natural Algorithms, PhD thesis, Politecnico di Milano, Italie, 1992.

[67] Ding, L. Chan, Classification of hyperspectral remote sensing images with support vector machines and particle swarm optimization, in: Proceedings of (ICIECS'09), pp. 1–52, 2009.

[68] S. Maldonado, R. Weber, A wrapper method for feature selection using SVM, Information sciences Elsveir, doi:10.1016/j.ins.2009.02.014, pp. 2208-2217, 2009..

[69] V.Sindhwani, P.Bhattacharya, and S. Rakshit. Information theoretic feature crediting in multiclass SVM. In Proceedings of the firrst SIAM International Conference on Data Mining, 2001.

[70] A. Al-Ani, A.Alsukker and R. N. Khushaba. Feature subset selection using differential evolution and a wheel based search strategy, Elsevier: Swarm and Evolutionary Computation 9, pp.15–26, 2013.