Effects of Using Filter Based Feature Selection on the Performance of Machine Learners Using Different Datasets

Mehnaz Khan [1] and **S. M. K. Quadri** [2]

Submitted in April, 2013; Accepted in August, 2013

Abstract - Data preprocessing is a very important task in machine learning applications. It includes the methods of data cleaning, normalization, integration, transformation, reduction, feature extraction and selection. Feature selection is the technique for selecting smaller feature subsets from the superset of original features/attributes in order to avoid irrelevant and additional features/attributes in the dataset and hence increases the accuracy rate of machine learning algorithms. However, the problem exists when the further removal of such features results in the decrease of the accuracy rate. Therefore, we need to find an optimal subset of features that is neither too large nor too small from the superset of original features. This paper reviews different feature selection methods- filter, wrapper and embedded, that help in selecting the optimal feature subsets. Further, the paper shows effects of feature selection on different machine learning algorithms- NaiveBayes, RandomForest and kNN). The results have shown different effects on the accuracy rates while selecting the features at different margins.

Index Terms - Data preprocessing, feature extraction, feature selection, dataset.

1. INTRODUCTION

In machine learning applications one of the most important tasks is data preprocessing [1]. The data that are collected for training in the machine learning tasks are not appropriate for the training purposes initially. In order to make the data useful for such applications, it needs to be processed. Processing involves methods for handling missing data [2] and methods for detecting and handling noise [3]. Data preprocessing is performed in order to prepare the data for input into machine learning and mining processes. This involves transforming the data for improving its quality and hence the performance of the machine learning algorithms, such as predictive accuracy and reducing the learning time. At the end of the data preprocessing stage, we get our final training set. One of the tasks of data preprocessing is feature selection in which only some of the features from the dataset are selected and used in the training process of the learning algorithm.

[1, 2]*Department of Computer Science, University of Kashmir, India*
E-mail: mhnzkhan6@gmail.com

In this process the aim is to find the optimal subset that increases the efficiency of the learning algorithm. Features in a dataset can be relevant i.e. the features that have influence on the output or irrelevant i.e. the features that have no effect on the output. Thus feature selection involves identifying the relevant features and using them in the machine learning application and ignoring the rest of the features with little or no predictive information.

The most important purpose of feature selection is to make a classifier more efficient by decreasing the size of the dataset. This is necessary for the classifiers that are costly to train e.g. NaiveBayes. The processing time and the cost of the classification systems are increased while their accuracy is decreased if irrelevant and additional features are used in the datasets being used for classification. Therefore, it is very important to develop the techniques for selecting smaller feature subsets. However, it has to be made sure that the subset which is selected is not so small that the accuracy rates are reduced and the results lack understandability. So it is very important that techniques must be developed that help to find an optimal subset of features from the superset of original features.

Feature selection comes with two approaches. One is called *forward selection* in which the process starts with no attributes/features which are then added one by one. At each step, the feature that decreases the error the most is added and the process continues until the addition of the features does not significantly decrease the error. Second approach is called *backward selection* in which the idea is to start with all the attributes/features and then remove them one by one. The feature to be removed at each step is the one that decreases the error the most, and the process is carried on until any further removal increases the error significantly.

In Section 3, different feature selection methods- filter, wrapper and embedded, that help in selecting the optimal feature subsets have been explained. Section 4 lists the basic steps that have been used for feature selection. Further, Section 5 gives the details of the experiment that was carried out using different machine learning algorithms on real data sets- Australian Credit Approval dataset from UCI Repository of Machine Learning Databases and Domain theories, Congressional Voting Records Dataset, and Adult Dataset. The optimal feature subsets achieved from the experiments have been explained in results and conclusion section.

2. RELATED WORK

This section presents the work done in the field of feature selection. A method of feature selection, called RELIEF has

been given that assigns a relevance weight to each attribute using instance based learning [4]. A book on feature selection has been given that includes all the feature selection methods that have been developed since 1970s and also gives a framework that helps in studying these methods [5]. Wrappers for feature subset selection have been developed in which an optimal feature subset is searched that is tailored to a particular learning algorithm and a particular training set [6]. The FOCUS algorithm has been designed for noise-free Boolean domains and it follows the MIN-FEATURES bias. It examines all feature subsets and selects the minimal subset of features that is sufficient to predict the class targets for all records in the training set [7]. Information gain and gain ratio are good examples of measuring the relevance of features for decision tree induction. They use the entropy measure to rank the features based on the information gained; the higher the gain the better the feature [8]. A feature selection model has been proposed using an instance-based algorithm, called RACE, as the induction engine, and leave-one-out cross-validation (LOOCV) as the subset evaluation function [9]. Emphasis has been laid on the issues of irrelevant features and the subset selection. It has been concluded that features that are selected should be dependent on the features and the target concept, as well as on the induction algorithm [10]. The forward and backward stepwise methods on the Calendar Apprentice domain have been designed, using the wrapper model and a variant of ID3 as the induction engine [11]. A method of feature selection for SVMs has been developed. The idea behind this method is to find those features which minimize bounds on the leave-one-out error. They have shown the method to be efficient as compared to some standard feature selection algorithms by testing on the datasets [12]. Twelve feature selection methods have been compared and a new feature selection metric called bi-normal separation (BNS) has been shown [13]. An introduction to variable and feature selection has been given that has suggested the use of a linear predictor e.g. a linear SVM and selection of variables in one of the two alternate ways. One is to use a variable ranking method using a correlation coefficient or mutual information and the other with a nested subset selection method that performs forward or backward selection [14]. A survey of feature selection methods for classification has been given [15]. A comparative study of feature selection methods in statistical learning of text categorization has been given that has evaluated document frequency (DF), information gain (IG), mutual information (MI) [16].

3. METHODS OF FEATURE SELECTION

Feature selection is regarded as a search problem in a space of feature subsets for which we need to specify a starting point, a strategy to traverse the space of subsets, an evaluation function and a stopping criterion [17]. There are three ways in which feature selection can be carried out. These are the filter, wrapper and embedded approaches. These methods differ in how they combine feature selection search with the construction of classification model [18, 23].

3.1 Filter Method

The filter approach selects a subset of the features that preserves as much as possible the relevant information found in the entire set of features. The methods that use the filter approach are independent of any particular algorithm as the function that they use for evaluation relies completely on properties of the data [24]. The relevance of the features is calculated by considering the intrinsic properties of the data. This involves the calculation of a feature relevance score and the features whose score is less are removed and only the remaining subset of features are used as input to the algorithm.
Some filter methods use correlation coefficients like that of Fisher's discriminant criterion. Other methods use mutual information or statistical tests. Initially the filter-based methods did not take into consideration the relations between features but calculated the relevance of each feature in isolation. However, now the filter methods take many criteria into consideration e.g. now the filter methods select features with minimum redundancy. The most important feature selection framework used by many filter methods is the minimum-redundancy-maximum relevance (MRMR) framework [28]. Further the filter methods can be univariate or multivariate. Univariate filter methods take into account only one feature's contribution to the class at a time, e.g. information gain, chi-square. These methods are computationally efficient and parallelable however they are likely to select low quality feature subsets. On the other hand, multivariate filter methods take the contribution of a set of features to the class variable into account at a time, e.g. correlation feature selection and fast correlation-based filter. These methods are computationally efficient and select high quality feature subsets than univariate filters.

3.1.1 Advantages of Filter Approach

Some of the advantages of filter approach are:

- Filter methods of feature selection can be easily scaled to very high-dimensional datasets.
- These methods perform very fast and are computationally simple.
- They are not dependent on any particular algorithm.
- In these methods, feature selection is to be carried out only once, and then different classifiers can be evaluated.
- These methods have better computational complexity as compared to the wrapper methods.

3.1.2 Disadvantages of Filter Approach

Filter approach has some drawbacks:

- These methods do not take into account the interaction with the classifier. In other words, this method separates the search in the feature subset space from the search in the hypothesis space.
- In this method each feature is measured separately and thus does not take into account the feature dependencies.
- Lack of feature dependencies results in the degraded performance as compared to other techniques.

However, this problem is solved by a number of multivariate filter techniques that involve feature dependencies to some extent.
Examples of this approach: Euclidian distance, t-test, information gain, correlation based feature selection, Markov blanket filter.

3.2 Wrapper Method

Filter methods use a function for evaluation that relies on the properties of data and thus is not dependent on any algorithm. On the other hand, wrapper methods make use of the inductive algorithm for calculating the value of a given subset. These methods take into account the biases of the algorithm and thus are considered to be a better alternative in supervised learning problems. Wrapper methods [27] include the model hypothesis search within the feature subset search. In this method, the subsets of features are selected by first defining a search process in the possible feature subsets space, followed by generating and evaluating various subsets of features. The subsets of features are evaluated by training a specific classification model. This makes the wrapper method algorithm specific. Then for searching the space of all possible feature subsets, this method wraps a search algorithm around the classification model. But the search process requires a number of executions which results in a high computational cost, especially when more extensive search strategies are used. The search methods are of two types: deterministic and randomized search algorithms. Wrapper methods make use of the classifier for scoring the subsets of features based on their predictive power. A number of wrapper methods have been developed that are based on SVM. Support Vector Machine Recursive Feature Elimination is a wrapper method that makes use of a backward feature elimination scheme for eliminating insignificant features from subsets of features. In this method the features are ranked on the basis of the amount of reduction in the function. The feature selected for elimination is the one with the lowest rank.

3.2.1 Advantages of Wrapper Method

Advantages of wrapper approaches are:

- These methods involve the interaction between feature subset search and model selection.
- Wrapper methods take into account feature dependencies.
- Implementing a wrapper method is quite easy and straightforward in supervised learning.

3.2.2 Disadvantages of Wrapper Method

Its drawbacks are:

- These methods have a higher risk of overfitting than filter techniques.
- Wrapper methods are computationally intensive.

Examples of this approach: Sequential forward selection, sequential backward elimination, beam search, genetic algorithms.

3.3 Embedded Method

Another type of feature selection technique is called embedded method. In this method the process of searching an optimal subset of features is included within the classifier construction, and is viewed as a search in the combined space of feature subsets and hypotheses. Embedded methods [25] are specific to a given learning algorithm like the wrapper methods. Advantage of these methods is that they include the interaction with the classification model like the wrapper methods and are less computationally intensive than wrapper methods. Examples of this approach are decision trees and artificial neural networks.
Examples of this approach: Decision trees, weighted NaiveBayes, feature selection using the weight vector of SVM.
Table 1 shows the comparison of various feature selection methods [26].

Feature Selection Methods	Advantages	Disadvantages
Univariate Filter	Classifier independence, scalability and fast speed.	Lack of feature dependencies and classifier interaction
Multivariate Filter	Includes feature dependencies. Classifier independence. Better computational complexity.	Lack of classifier interaction, less scalable and slower.
Deterministic Wrapper	Includes classifier interaction, presence of feature dependencies	Classifier dependence, risk of overfitting.
Randomized Wrapper	Includes classifier interaction, presence of feature dependencies	Classifier dependence, high risk of overfitting, computationally intensive.
Embedded	Better computational complexity, includes classifier interaction	Classifier dependence.

Table 1: Comparison of Feature Selection methods

4. STEPS USED IN FEATURE SELECTION

This section discusses the steps followed in selecting the subset of features using filter approach. The steps used are:

- Initialize the learner.
- Load the dataset.
- Create the classifiers by training the learner on the dataset.
- Compute the relevance of the features.
- Set some margin, say *m*, and remove all those features for which relevance is less than *m.*

- Only the features whose relevance is greater than *m* are used for classification.
- Finally, use the learner on both the datasets and compare the accuracy.

5. EFFECTS OF FEATURE SELECTION ON VARIOUS DATASETS

The above mentioned steps were implemented on various machine learning algorithms (RandomForest, NaiveBayes and kNN) using different datasets used in our experiments. The experiments were carried out in Python programming using python machine learning tool. The datasets that were used in the experiments are the Australian Credit Approval dataset [19] from UCI Repository of Machine Learning Databases and Domain theories, Congressional Voting Records Dataset [20], and Adult Dataset [21]. The Credit dataset has already been used in the evaluation of the machine learning techniques earlier in our work [22].

5.1 Experiment

Different experiments were performed using different machine learners on the three datasets mentioned above. The machine learning algorithms used in the experiments are RandomForest, NaiveBayes and kNN. First the performance of the learning algorithms without feature selection has been shown on all the three datasets. After that performance of the learners has been checked by applying feature selection at margins 0.01, 0.02, 0.03 and 0.04. At different margins, it was observed that a number of irrelevant attributes got discarded depending on the calculated relevance of the attributes and only the relevant ones were used in the learning process.

Figure 1 shows the performance of machine learning algorithms on the Adult Dataset without using feature selection. Figure 2 shows the performance of machine learning algorithms on the same dataset using feature selection at margin 0.01. Figure 3 shows the performance of machine learning algorithms on the same dataset using feature selection at margin 0.03. Figure 4 shows the performance of machine learning algorithms on the same dataset using feature selection at margin 0.04.

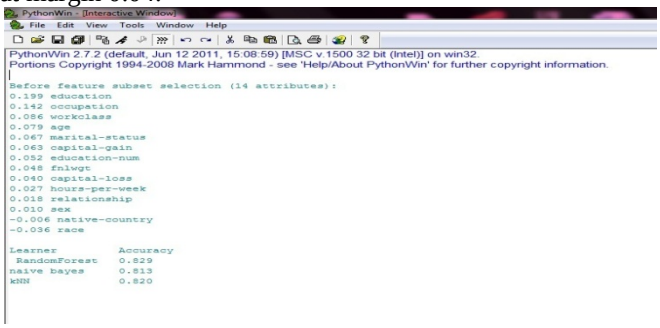

Figure 1: Results of Adult Dataset without feature selection

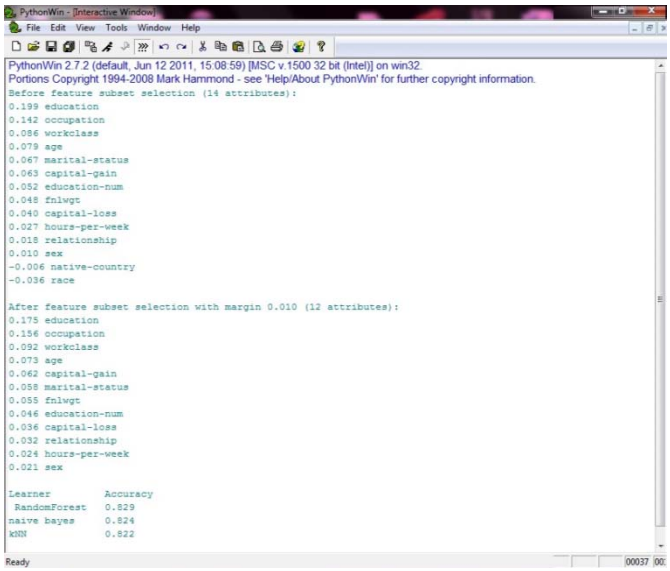

Figure 2: Results of Adult Dataset at margin 0.01

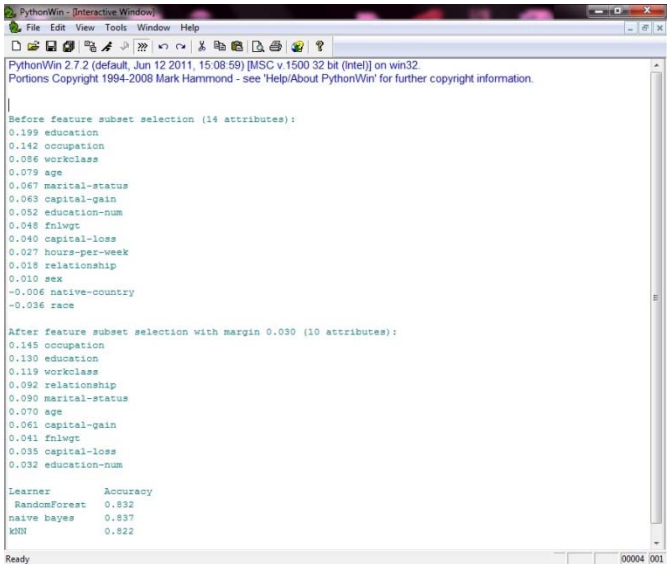

Figure 3: Results of Adult Dataset at margin 0.03

Table 2 shows the performance of the machine learners on the Adults Dataset in terms of accuracy, first without feature selection (FS) and then with feature selection at different margins. Table 3 shows the performance of the machine learners on the Credit Dataset in terms of accuracy, first without feature selection and then with feature selection at different margins. Table 4 shows the performance of the machine learners on the Voting Dataset in the same way.

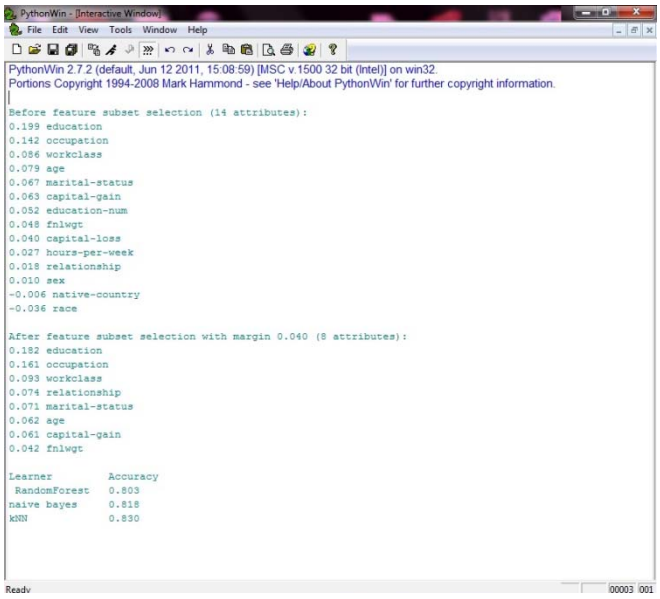

Figure 4: Results of Adult Dataset at margin 0.04

Learners	Accuracy				
	Without FS	**With FS (0.01)**	**With FS (0.02)**	**With FS (0.03)**	**With FS (0.04)**
RandomForest	0.829	0.829	0.829	0.832	0.803
NaiveBayes	0.813	0.824	0.829	0.837	0.818
kNN	0.820	0.822	0.822	0.822	0.830

Table 2: Results of Adult Dataset at different margins

Learners	Accuracy				
	Without FS	**With FS (0.01)**	**With FS (0.02)**	**With FS (0.03)**	**With FS (0.04)**
RandomForest	0.845	0.852	0.838	0.837	0.850
NaiveBayes	0.864	0.864	0.858	0.851	0.830
kNN	0.831	0.835	0.831	0.851	0.831

Table 3: Results of Credit Dataset at different margins

Learners	Accuracy				
	Without FS	**With FS (0.01)**	**With FS (0.02)**	**With FS (0.03)**	**With FS (0.04)**
RandomForest	0.956	0.959	0.959	0.954	0.954
NaiveBayes	0.903	0.915	0.915	0.915	0.915
kNN	0.936	0.936	0.929	0.936	0.936

Table 4: Results of Voting Dataset at different margins

5.2 Results and Discussions

For all the datasets the learning algorithms have shown an increase in accuracy after feature selection. In some cases, the efficiency of learning algorithms after feature selection remained same as it was before feature selection depicting the fact that the discarded features were irrelevant and contributed nothing towards their performance as there was no change in accuracy even after discarding them. And hence were not needed. However, in some cases the learning algorithms show an increase in efficiency after feature selection. However, it increases only up to a certain limit. After that accuracy starts to decrease if feature selection is continued as more and more features are being discarded. Figure 1 and Table 2 show the results for Adult Dataset. Initially when no feature selection process is carried out, all the attributes of the Adult dataset (i.e. 14 attributes) are used in the learning process and the efficiencies of learners are 0.829 for RandomForest, 0.813 for NaiveBayes and 0.820 for kNN. After that when feature selection is carried out at margin 0.01, two of its attributes are discarded as their relevance is below the margin and thus only 12 attributes are used in the learning process. At this margin there is an increase in the accuracy of the learners e.g. NaiveBayes- 0.824 and kNN- 0.822 or it remains constant e.g. RandomForest- 0.829. However, at margin 0.04, four of its attributes are discarded and only eight attributes are used in the learning process. At this stage the accuracy starts decreasing. This shows that we have to find an optimal subset of features for a dataset. Similar effects have been shown with other two datasets as well wherein feature selection shows an increase in the accuracy of all the three machine learners up to a certain limit after which accuracy starts decreasing. Table 5 shows the number of attributes of all the datasets before feature selection (i.e. original number) and the number of attributes that were used in the learning process after feature selection.

Datasets	Number of Attributes at different margins				
	Before FS	**0.01**	**0.02**	**0.03**	**0.04**
Adult	14	12	12	10	8
Credit	15	11	6	5	4
Voting	16	14	14	13	13

Table 5: Number of attributes at different margins

The results in Table 5 show that in case of Adult dataset only 10 attributes among 14 were relevant because the learners showed better or similar performance even after discarding these 4 attributes. However, using only 8 attributes decreases the accuracy. In case of Credit dataset 11 out of 15 attributes were relevant and in case of voting dataset only 13 among 16 were relevant as it showed better or similar performance after discarding 3 of its attributes.

CONCLUSION

Different machine learning algorithms- NaiveBayes, RandomForest and kNN were used for evaluation on real data sets before and after feature selection taking into consideration user defined limits or margins (i.e. 0.01, 0.02, 0.03 and 0.04). The results have shown different effects on the accuracy rates while selecting the features at different margins. As we realized from the experiment that after increasing the margin beyond certain limit the performance starts degrading. Hence it is necessary to find an optimal subset of features for each dataset.

REFERENCES

[1]. Zhang S., Zhang C., Yang Q. (2002), "Data Preparation for Data Mining", *Applied Artificial Intelligence*, Vol. 17, pp.375-381.

[2]. Batista G. and Monard M.C. (2003), "An Analysis of Four Missing Data Treatment Methods for Supervised Learning", *Applied Artificial Intelligence*, Vol. 17, pp.519-533.

[3]. Hodge V. and Austin J. (2004), "A Survey of Outlier Detection Methodologies", *Artificial Intelligence Review*, Vol. 22, Issue 2, pp. 85-126.

[4]. Kira K. and Rendell L. (1992), "The Feature Selection Problem: Traditional Method and a New Algorithm", in *Proceedings of the Tenth National Conference on Artificial Intelligence*, pp. 129-134. MIT Press,

[5]. Liu H. and Motoda H. (1998), "Feature Extraction, Construction and Selection: A Data Mining Perspective". *Kluwer Academic Publishers*, Boston/Dordrecht/London, Boston.

[6]. Kohavi R. and John G. (1997), "Wrappers for Feature Subset Selection". *Artificial Intelligence'97*, pp. 273-324.

[7]. Almuallim H. and Dietterich T. (1991), "Learning with Many Irrelevant Features", *Proceedings of the Ninth National Conference on Artificial Intelligence*, pp. 547-552. MIT Press.

[8]. Quinlan J. (1993), "C4.5: Programs for Machine Learning", *Morgan Kaufmann*, San Mateo.

[9]. Moore W. and Lee S. (1994), "Efficient Algorithms for Minimizing Cross Validation Error", *Machine Learning: Proceedings of the Eleventh International Conference.*

[10]. John G., Kohavi R. and Pfleger K. (1994), "Irrelevant Features and Subset Selection Problem", *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 121- 129. Morgan Kaufmann Publishers.

[11]. Caruana R. and Freitag D. (1994), "Greedy Attribute Selection", *Machine Learning: Proceedings of the Eleventh International Conference*, W. Cohen and H. Hirsh (eds). Morgan Kaufmann.

[12]. Weston J., Mukherjee S., Chapelle O., Pontil M., Poggio T. and Vapnik V. (2001), "Feature selection for SVMs", *Advances in Neural Information Processing Systems* 13. Cambride, MA: The MIT Press, pp. 668–674.

[13]. Forman G. (2003), "An extensive empirical study of feature selection metrics for text classification", *Journal of Machine Learning Research*, Vol. 3, pp. 1289–1305.

[14]. Guyon I. and Andr´e E. (2003), "An introduction to variable and feature selection", *Journal of Machine Learning Research*, Vol. 3, pp. 1157–1182.

[15]. Dash M. and Liu H. (1997), "Feature selection for classification", *Intelligent Data Analysis*, Vol. 1(1–4), pp. 131–156.

[16]. Yang Y. and Jan O. P. (1997), "A comparative study of feature selection in text categorization", *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 412–420.

[17]. Talavera, L. (2005), "An evaluation of filter and wrapper methods for feature selection in categorical clustering", *Advances in Intelligent Data Analysis VI Springer-Verlag*, pp. 440-451.

[18]. Saeys Y., Inza I. and Larranaga P. (2007), "A review of feature selection techniques in bioinformatics", *Bioinformatics* Vol. 23 No. 19, pp. 2507–2517.

[19]. Australian Credit Approval dataset http://www.hakank.org/weka/credit.arff from UCI Repository of Machine Learning Databases and Domain theories (http://archive.ics.uci.edu/ml/datasets.html). Accessed on 20-04-2013.

[20]. Congressional Voting Records Dataset (http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records. Accessed on 20-04-2013.

[21]. Adult Dataset http://archive.ics.uci.edu/ml/datasets/Adult. Accessed on 20-04-2013.

[22]. Khan M. and Quadri S.M.K. (2012), "Evaluating Various Learning Techniques For Efficiency", *International Journal of Engineering and Advanced Technology (IJEAT)*, Vol. 2, Issue 2, pp. 326-331.

[23]. Zenglin X., Rong J., Jieping Y., Michael R. L. and Irwin K. (2009), "Discriminative semi-supervised feature selection via manifold regularization", *IJCAI' 09: Proceedings of the 21th International Joint Conference on Artificial Intelligence*.

[24]. Brown G., Pocock A., Zhao M.J. and Lujan M. (2012). "Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection", *Journal of Machine Learning Research (JMLR).*

[25]. Canedo V.B., Maroño N.S. and Betanzos A.A. (2013), "A review of feature selection methods on synthetic data", *Springer-Verlag*, *Knowl Inf Syst* Vol. 34, pp. 483–519.

[26]. Ladha L. and Deepa T. (2011), "Feature Selection Methods and Algorithms", *International Journal on Computer Science and Engineering (IJCSE)* Vol. 3 No. 5, pp. 1787-1797.

[27]. Yang P., Liu W., Zhou B.B., Chawla S. and Zomaya A.Y. (2013), "Ensemble-Based Wrapper Methods for Feature Selection and Class Imbalance Learning", *Springer-Verlag* pp. 544-555.

[28]. Auffarth B., Lopez M. and Cerquides J. (2010), "Comparison of Redundancy and Relevance Measures for Feature Selection in Tissue Classification of CT images", *Advances in Data Mining Applications and Theoretical Aspects, Springer* pp. 248-262.