International Journal of Database Theory and Application
Vol.9, No.9 (2016), pp.75-82
http://dx.doi.org/10.14257/ijdta.2016.9.9.07

# Performance Evaluation of Feature Selection Methods on Large Dimensional Databases

Y. Leela Sandhya Rani[1], V. Sucharita[1], Debnath Bhattacharyya[2] and Hye-Jin Kim[3]

*1Department of Computer Science and Engineering,*
*KL University, Vaddeswaram, AP, 522502, India*
*2Department of Computer Science and Engineering,*
*Vignan's Institute of Information Technology,*
*Duvvada, Visakhapatnam, India*
*3Sungshin W. University, 2, Bomun-ro 34da-gil,*
*Seongbuk-gu, Seoul*
*lsraniyalamarthi18@gmail.com, drvsucharita@kluniversity.in,*
*debnathb@gmail.com, hyejinaa@daum.net*
*(Corresponding Author)*

***Abstract***

*Data mining retrieves knowledge information from larger amounts of data. Clustering is an assemble of similar objects in to one class and dissimilar objects in to another class. When designing clustering ensemble on large dimensional data space, both time and space requirements for processing may be overinflated. This tends to impose feature selection methods to remove redundant features and handle the noise data. There are filter, wrapper and hybrid methods in feature selection. This paper shows a tour on types of feature selection techniques and numbers of experiments are conducted to compare feature selection techniques using different datasets with R tool, which gives better technique for clustering ensemble design.*

***Keywords****: Feature selection techniques, filter method, wrapper method, hybrid method and performance of feature selection techniques*

## 1. Introduction

Data Mining is a process of analyzing the data and requires only the necessary information for any type of applications [4]. Different tasks are specified for mining data from large amounts of data. One of main task in mining is clustering, is assemble the objects from same cluster having similar properties with other clusters having dissimilar properties. Clustering ensemble is the extension of clustering which apply clustering on set of subsets from the original data and then add all the results into one clustering using some consensus function.

Detecting relevant features, removing irrelevant features, redundant data, noisy data [2] is very crucial factor in mining the data. Working on large dimensional data space is crucial task, the algorithms applied on these data spaces the performance is almost poor.

To process high dimensional databases it is necessary to reduce feature space. Two approaches are used to reduce the dimensions:

1. Feature selection
2. Dimensionality reduction

In feature selection it selects a set of the features from the original space. But in dimensionality reduction a transformation process is applied on features of original space and produce less features from original space. The difference between feature selection

ISSN: 2005-4270 IJDTA

and dimensionality reduction is a subset of features is retained in feature selection but modified reduced features set is in dimensionality reduction.

Mainly the feature selection is used to retrieve selected feature set from the original feature space. Feature selection is used in many areas such as machine learning, classification of gene data in biology clustering ensemble text mining. Feature selection increases the accuracy of relevant features and reduces the overhead of selecting features. If we remove the irrelevant features automatically reduce the size of dataset and it is easy further processing.

In feature selection we have several aims [3]:

1. The execution time and memory space required to run our algorithms should be reduced.
2. To improve classifiers by removing noisy or irrelevant features.
3. To identify which features may be relevant to a particular problem.

From the three kinds of feature selection techniques, separate algorithms are there for each one. In the filter method Pearson correlation coefficient, relief algorithms are there. In wrapper method recursive feature elimination is there. The accuracy for this method is shown in tables and performance is given in the graph.

The advances of this process are increasing learning accuracy, execution speed up for data mining algorithms and fit better model.

The overview of this paper is twofold. First we illustrate all the feature selection techniques in detail and second we adopted different experiments for feature selection techniques and to compare the accuracy and performance evaluation of the algorithms using R tool.

The remaining paper is organized as follows. Section 2 describes previous work related to feature selection techniques Section 3 gives a tour on feature selection methods Section 4 presents experimental results in some tables and graphs Section 5 gives the conclusion.

## 2. Related Work

Feature selection is to get selective features from the original data set which avoids the redundant features and handle the noisy data. Feature selection is a common technique for selective data analysis, which is used in many fields, including machine learning, clustering ensemble, pattern recognition, image processing and scientific applications such as to classify fusion data in nuclear physics [1].

Feature selection is used in many areas of research for example in the area of designing clustering ensemble there is a requirement of feature selection. Feature selection has several methods. To choose a particular method it is important to analyze the accuracy and performance evaluation of feature selection methods.

The first method is filter method, in that several algorithms are there Pearson correlation Coefficient, Mutual information, Relief, Ensemble with data permutation, Ensemble of methods. In wrapper approach greedy forward search, exhaustive search algorithms are specified. In hybrid method ranked forward search, refined exhaustive search are specified. The author Matthew Shardlow works with feature selection methods in large data sets and concluded that wrapper methods are more accurate than filter methods [3].

The three algorithms, relief, fast clustering algorithm and k-means clustering are specified in the table that include number of the original features and selected features [4]. The author Pinar Yildirim illustrates the feature selection methods and evaluates the performance using four classification methods [5].

The author proposed relief method and hybrid method and identified that hybrid method outperforms the relief method regarding misclassification error rate and running

 

time [6]. The authors implement wrapper approach based on genetic algorithm and case based reasoning and results shows that the model is better than other models [7].

From these articles we have an idea to identify which feature selection methods are better for designing clustering ensemble. And we tabulate the performance evaluation and accuracy of the algorithms.

## 3. Feature-Selection Methods

Feature selection methods are used in wide variety of fields such as image processing to extract features from the image, biology to processing gene features, machine learning and many more. Feature extraction means to select relevant features to reduce the processing time. Depending upon its design the feature selection methods are differentiated in to three methods.

1. Filter Methods.
2. Wrapper Methods.
3. Hybrid Methods.

### 3.1. Filter Methods

Filter methods carry out the feature selection process as a initial step with no induction algorithm. The filter method cannot use any classifier and it uses distance between the classes or how one feature will statistically depends on another feature. The filter methods are faster because it depends only on the induction algorithm and no classifier is used but in terms of accuracy wrapper methods are efficient. However, it leads to select a feature set with a high number of features and a threshold is required to choose a feature subset. There are different types of algorithms are there.

#### 3.1.1. Pearson Correlation Coefficient

In large dimensional databases features are highly correlated with each other. Algorithms applied on these data sets they perform poor. so if we remove highly correlated features the performance of algorithms is better. The Pearson correlation coefficient measures the capacity between variables and relationships. To compare two features, it is a good idea to conduct a Pearson correlation coefficient value to determine how well that relationship is between those two features. In R using Caret package we can use find Correlation function which will analyze a correlation matrix of features that can report the features to be removed.

Pearson's correlation coefficient when applied to a sample is represented by the letter r and may be referred to as the sample correlation coefficient or the sample Pearson correlation coefficient. We can obtain a formula for r by substituting estimates of the co-variances and variances based on a sample into the formula above. So if we have one dataset $\{x_1,...,x_n\}$ containing n values and another dataset $\{y_1,...,y_n\}$ containing n values then that formula for r is:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} \quad (1)$$

where $\bar{x}$ and $\bar{y}$ are the averages of the sample.

#### 3.1.2. Relief

This method [3] separate the features of randomly selected instances *i.e.,* to select each instance of the data set having both nearest same class instance and nearest opposite class instance. Then these are used to calculate a weighting for each feature which is iteratively updated with each chosen data point. This method can be highly useful when there is a large amount of data. Time complexity is not an issue as a constant number of trials is

performed. This means that the relief algorithm is speed than other methods which require all the data to be taken into account.

#### 3.1.3. Information Gain

Information gain is the variation between two probability distributions. It assesses a feature X by calculating the amount of information gained with respect to the class variable Y, defined as follows:

$$I(X) = H(P(Y)-H(P(Y/X)) \quad (2)$$

Specifically, it evaluates the difference the marginal distribution of observable Y assuming that it is independent of feature X(P(Y)) and the conditional distribution of Y assuming that it is dependent of X (P(Y/X)). If X is not differentially expressed, Y will be independent of X, thus X will have small information gain value, and vice versa [5].

### 3.2. Wrapper Methods

Wrapper methods are based on classifiers. It is a search problem, where different sets are prepared, evaluated and compared to other sets. A predictive model us used to assess a set of features and assign a threshold based on model accuracy. Depending upon the accuracy for the selected features we finalize the feature subset.

Different wrapper approaches are there including best-first search, hill climbing algorithm, forward and back ward passes to add and remove features.

#### 3.2.1. Recursive Feature Elimination

First, the algorithm fits the model to all features. Each feature is ranked using its importance to the model. Let S be a series of ordered numbers which are candidate values for the number of features to retain ($S_1 > S_2$, ...). At each iteration of feature selection, the Si top ranked features are remained in the set, the model is trained and performance is evaluated. The value of $S_i$ with the best performance is determined and the top $S_i$ predictors are used to fit the final model. Recursively eliminate the feature depending upon the accuracy obtained from the inclusion of that feature.

## 4. Results and Discussions

### 4.1. Tool Description

R is mostly used tool for statistical analysis, machine learning and data mining. It was developed at Bell Laboratories by John Chambers and colleagues. R has tools for linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and graphical techniques, and is highly extensible. R become very popular with the academic and industrial researchers, and is also widely used for scientific purposes.

### 4.2. Data Sets

Four data sets are used for to select the features. These are taken from the R packages and meant for classification problems. Table 1 shows the specification of data for testing purposes, the dataset is described by the type of the data being used, the types of attributes, whether they are real, categorical, integer, Factor the number of instances stored within the data set, the number of attributes that describe dataset.

These data sets are selected because they has large set of features mainly used for classification problems.

 

**Table 1. Description of the Datasets**

| Data sets | Feature Type | Instances | Features |
|---|---|---|---|
| PimaIndiansDiabetes | Numeric, Factor | 768 | 9 |
| Vehicle | Numeric | 846 | 19 |
| BloodBrain | Numeric | 208 | 134 |
| mdrr | Numeric, Factor | 528 | 342 |

### 4.3. Feature Selection Using Filter Methods

Without classifier Filter methods use only statistical measures. These methods implementation and workout is fast when compared to other methods.

#### 4.3.1. Pearson Correlation Coefficient

Highly redundant features correlated more. Removing highly correlated features enhance the classification accuracy. In R tool the find correlation function used to retrieve the highly correlated features. Using all the four tables mentioned in Table1 we applied the function and tabulated the results in Table 2.

**Table 2. Highly Correlated Features**

| Data sets | Instances | Features | Highly corre-lated Features |
|---|---|---|---|
| PimaIndiansDiabetes | 768 | 9 | 1(cutoff=0.5) |
| Vehicle | 846 | 19 | 13(cutoff=0.5) |
| BloodBrain | 208 | 134 | 75(cutoff=0.7) |
| mdrr | 528 | 342 | 288(cutoff=0.7) |

#### 4.3.2. Relief

The relief method selects each instance both nearest same class data point and nearest opposite class data point. These are then used to calculate a weighting for each feature which is iteratively updated with each chosen data point. Using this weight a feature importance can be calculated.

Using the relief function in R tool we selected the features based on the feature importance. We graphically represented the important features for four data sets.

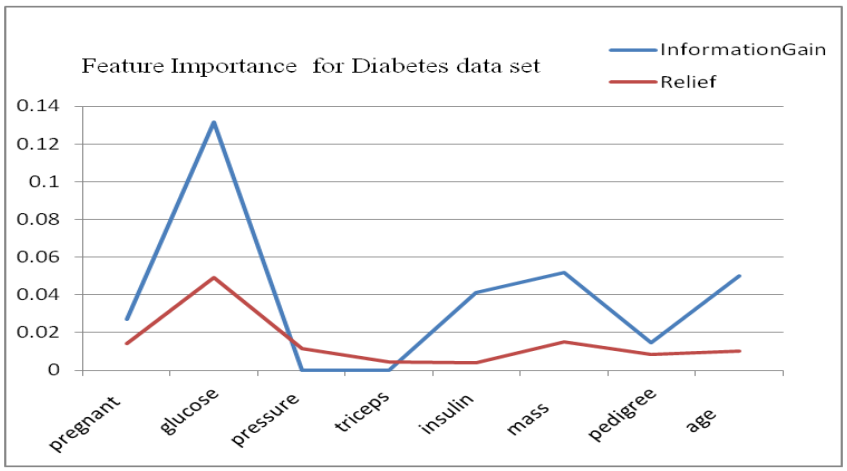


**Figure 1. Feature Importance Using Relief and Information Gain on Diabetes Dataset**

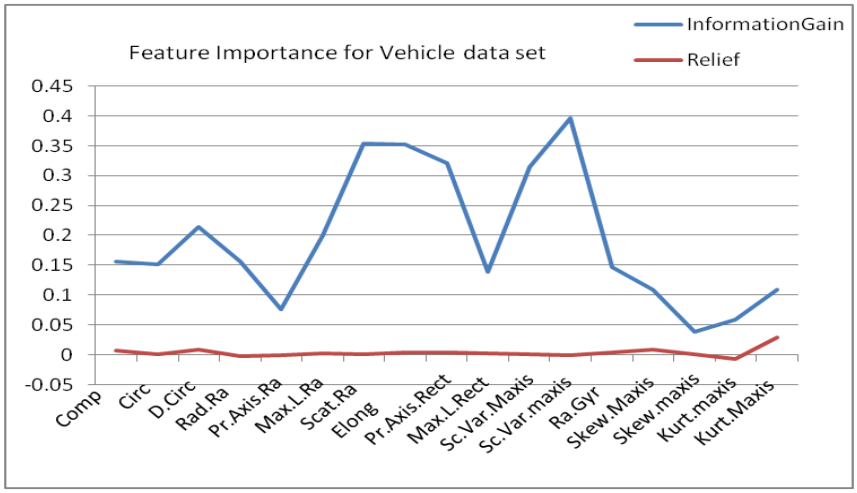


**Figure 2. Feature Importance Using Relief and Information Gain on Vehicle Dataset**

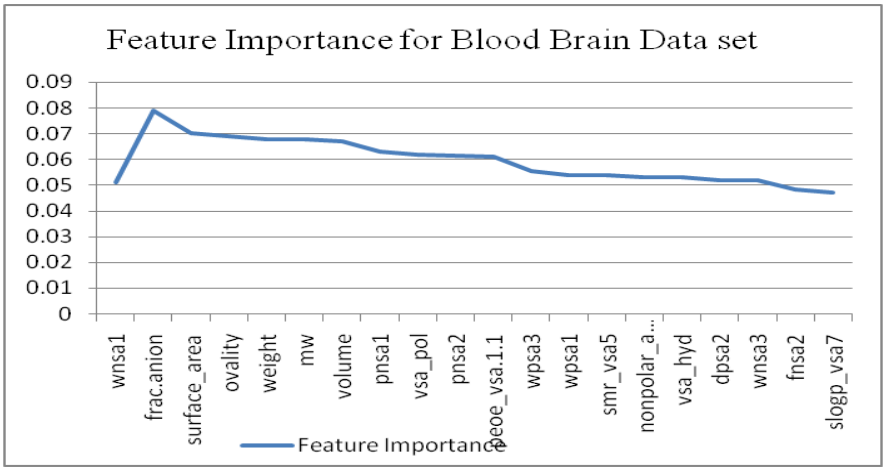


**Figure 3. Feature Importance Using Relief on Blood Brain Dataset**

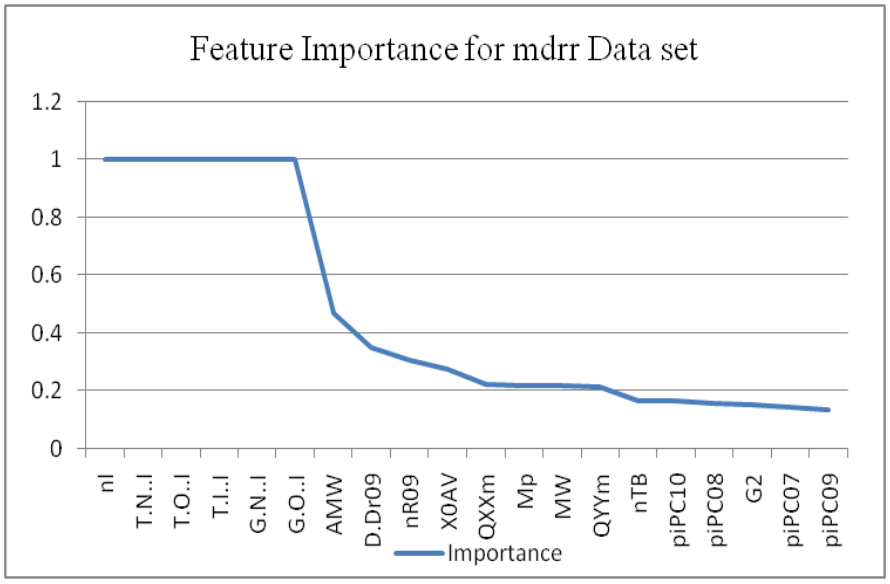


**Figure 4. Feature Importance Using Relief on mdrr Data set**

### 4.3.3. Information Gain

We selected set of important features using information gain function in R tool for two different data sets one is Diabetes data set and another one is vehicle data set.

### 4.4. Feature Selection Using Wrapper Methods

In wrapper method a classifier is used to train the data and obtain set of features. These methods are more accurate than that of filter methods. We can also evaluate classification accuracy for the selected features definitely it is better than classification accuracy for all the features.

### 4.4.1. Recursive Feature Elimination

First train the classifier with all features then remove certain features depending upon its accuracy. The steps can be done repeatedly until accurate features left in the set.

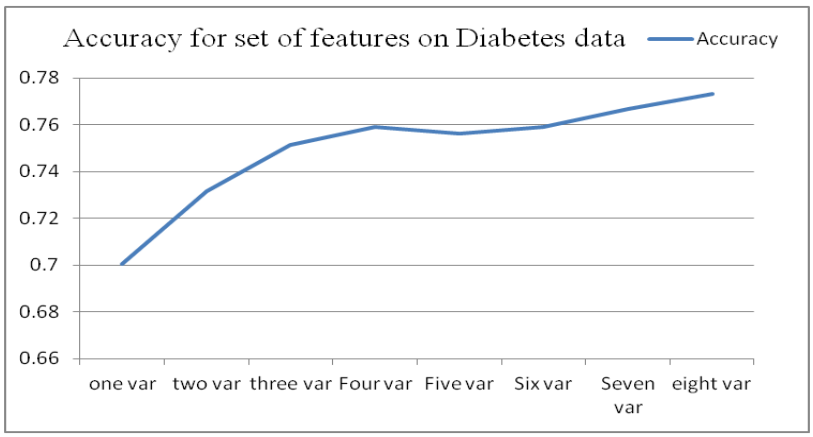


**Figure 5. Accuracy for Features Using RFE on Diabetes Dataset**

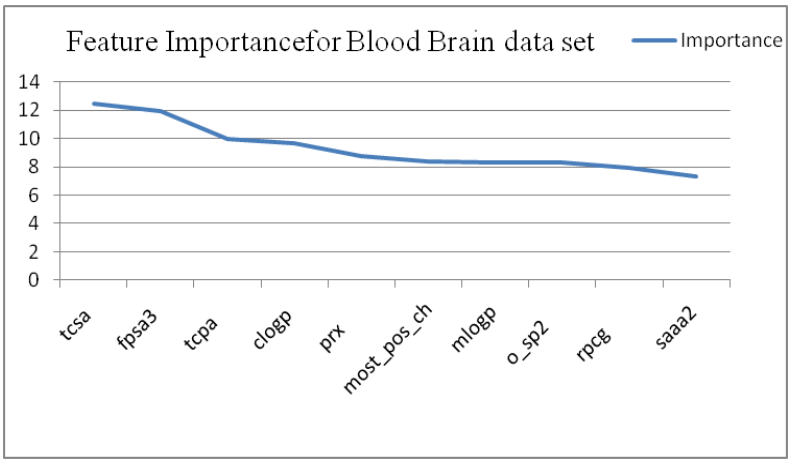


**Figure 6. Feature Importance Using RFE on Blood Brain Dataset**

In PimaIndiansDiabetes dataset eight features are there, using RFE feature selection method we obtained top five important features. The classification accuracy for five features and all features is tabulated in Table 3.

**Table 3. Accuracy of Feature Selection**

| Data sets | Features | Accuracy | Kappa |
|---|---|---|---|
| Diabetes | 5 | 0.7760417 | 0.4762751 |
| Diabetes | 8 | 0.66145833 | 0.04703726 |

## 5. Conclusion

Feature selection methods are widely used in image processing, clustering ensemble, due to its capacity to improve the classification accuracy and reduce the redundant features. In this paper we presented filter wrapper methods most widely used in feature selection areas. We tested filter wrapper methods on four different data sets and identified that the execution is fast for filter methods than wrapper methods. The wrapper methods are more accurate than filter methods. Further we want to proceed to test more wrapper and hybrid methods and its performance.

## References

[1] C. P. Erick, S. Newsam and C. Kamath, "Feature selection in scientific applications", Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, Seattle, WA, United States, **(2004)**, pp. 788-793.
[2] S. V. Jadhav and V. Pinki, "A Survey on Feature Selection Methods for High Dimensional Data", International Journal on Recent and Innovation Trends in Computing and Communication, IJRITTC, vol. 4, no. 1, **(2016)**, pp. 83-86
[3] S. Matthew, "An analysis of feature selection techniques", The University of Manchester, **(2016)**, pp. 1-7.
[4] P. Karthika, S. N. Roopa and S. Monisha, "Comparison of Performance of Feature Selection Methods", International Journal of Advanced Research in Computer and Communication Engineering, vol. 5, no. 3, **(2016)**, pp. 239-241.
[5] Y. Pinar, "Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease", International Journal of Machine Learning and Computing, vol. 5, no. 4, **(2015)**, pp. 258-263.
[6] D. Luis and E. Acuna, "Feature selection based on a data quality measure", Proceedings of the World Congress on Engineering, WCE, London, U.K, vol. 2, **(2008)**, pp. 1095-1099.
[7] D. Mohammad, A. A. Liaei, and M. Hosseini, "Feature selection for breast cancer diagnosis: a case-based wrapper approach", World Academy of Science, Engineering and Technology, International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering, vol. 5, no. 5, pp. 220-223.

 