Integrated Intelligent Research (IIR)
International Journal of Computing Algorithm
Volume: 08, Issue: 01, June 2019, Page No.8-13
ISSN: 2278-2397

Significance of Feature Selection Impacting Good Clusters

S.Anitha[1], Mary Metilda[2]
[1]Research Scholar, Bharathiar University,Coimbatore
[2]Asst. Prof, Queen Marys College, Chennai
Email: anitasenthil@gmail.com, metilda_dgvc@yahoo.co.in

Abstract - Feature selection is one of the recent techniques in data mining. Clustering is an unsupervised learning method in knowledge discovery world. Dimensionality reduction poses electing the significant data from the wider variety of data elements. "Curse of dimensionality "is a challenging issue of present researches which produces wrong outcomes during clustering process. In this paper, cluster based outlier detection method is applied with various multivariate datasets. Before clustering the datasets, the feature selection method has been implemented for selecting significant datasets from the entire training attributes. Feature selection plays an essential role in cluster accuracy for obtaining the dissimilar and dissimilar datasets among the data instances. In this research work, the proposed system is compared with the various correlation based feature selection algorithms and their experimental results are depicted.

Keywords: Data Clusters, Correlation Feature Selection, Feature Selection, Genetic Algorithm, Multivariate datasets

I. INTRODUCTION

In recent, enormous electronic data are being utilized and sustained by corporate across the world. The wider ranges of data are impractical for human analyst at high speed. The traditional statistical methods are inadequate to analyze these types of data. So the researchers have to look for an alternative and hence develop advanced data mining tools with knowledge that will help in the decision-making process. The major point of the data mining process is to "revolve data into knowledge". High dimensionality is one of the fast growing problems in Computer Science, because of the increased requirement for tools that help in the analysis of huge amounts of data. The major aspire of a feature selection approach in data mining is to eliminate unrelated attributes from a problematic area. Particularly in health care domain, appropriate clinical diagnosis should be done before sending data for laboratory confirmation to avoid false declaration of diseases. In this research, the medical datasets are chosen for analysing in terms of size of the dataset and number of attribute of the data with class labels.

Huge amount of data points implemented in algorithm was unable to terminate in a stipulated amount of time [1]. Reducing irrelevant attributes increases the performance of clustering process before identifying outliers [2]. Feature selection methodology is used to eliminate insignificant features and to construct a minimum attribute subset using the data alone, without the need of any supplementary information [3]. It has adopted Correlation coefficient based feature selection (CFS) [4]. It is combined with different types of searching methods to determine the important attribute subsets [5], [6]. In the proposed research, Genetic algorithm based feature selection is implemented for retrieving relevant subsets for further process. It is used to estimate correlation between subset of attributes, class variables and inter correlation between the attributes as well [7], [8]. The selected features construct a classification form. It utilizes the advantages of genetic algorithm combining with correlation feature selection to progress the classification task. The proposed method is evaluated by conducting experiments on PIMA Indian diabetes dataset taken from UCI Machine Learning Repository. PIMA dataset consists of 768 instances and 9 real valued input features [9]. Out of which 268 points are with diabetes and 500 are without diabetes disease with 376 data points contain missing values. Pima Indian Diabetes multivariate datasets includes 8 attributes with a class attribute. In proposed technique, first phase is called Genetic Based

International Journal of Computing Algorithm
Volume: 08, Issue: 01, June 2019, Page No.8-13
ISSN: 2278-2397

Feature Selection (GBFS) is evaluated on various data mining classifiers such as Naive Bayes (NB), Multilayer Perceptron (MLP), Sequential Minimal Optimization (SMO), Radial Basis Function Network (RBFN), and random forest. Further section describe the new feature selection technique using genetic based features selection (GBFS) and Correlation Feature Selection (CFS) with succeeding analytical discussions. This paper has organized as follows: section II describes the genetic algorithm. Section III discusses results and discussions of the research. Section IV analysis various feature selection methods for evaluating datasets Section V concludes the proposed method

II. GENETIC BASED FEATURE SELECTION

Genetic algorithm consisting of strings that are represented as parameters and the set of parameters represent population. [10]. According to Darwin's theory of ''Survival of Fitness'', the strings are randomly selected then induced to crossover and mutation. The genetic algorithm is used to provide a new subpopulation or off-spring. The process of selecting the objects from the population continues until required subsets are reached based on the termination criterion or the chosen set of solution attains the fixed number of generation [11]. It is used to classify or cluster the labelled or unlabeled dataset.

Algorithm-1 : Algorithm GBFS( )
// Input: X: {xi ..... xn} be a set of training data points.
// P: Initial Population,x1, y1: new population members
// Output: Significant subset of Data Points S

Step: 1 Select data points as X
Step: 2 Apply CFS for the dataset to obtain reduction in the features Step: 3 Apply genetic search strategy for calculating fitness function. Do{

(1) P=initial set of population p strings
(2) Begin (randomly) generating an initial **population** P.
(3) If the solution is satisfied then terminate else jump to next step
(4) Evaluate **fitness** value.
(5) Initialize number of generation.
(6) While number generation * 2 ≤ termination condition; do
(7) **Select** all the genetic solutions which can transmit to next generation (x1,y1)
(8) Increment number of generation
(9) Perform **crossover** operation up to until 50% of bits are crossed.
(10) End while.
(11) If the solution is efficient then apply **mutation.**
(12) Genetic algorithm searches best solution from a large set of Solutions
(13) Go to and repeat step (2). Return Optimal Subset of features S

The feature select is most familiar concept in data mining and it is also called as subset selection or attributes (variable) selection. It is used to remove the irrelevant and noisy variables for brief data representation. In this framework, genetic algorithm can be suitably employed to identify the subsets of relevant features. The reduction of features is based on data dependencies and is calculated by a fitness function. Genetic based Feature Selection (GBFS) is used to produce significant subset and forecast both diagnosis and prognosis by comparing several data mining classifiers. In this approach, pre processing stages are included to eliminate noise and inconsistent data in the dataset and formulate the dataset fit for promoting further processing. As the first part, the missing values in the datasets are replaced by mean value of the individual variable. The data mining classifiers do not work well with numeric features. It is necessary to transfer or discretize the numeric attributes in the dataset. An instance filter discretizes a series of numeric attributes in the dataset into nominal attributes.

International Journal of Computing Algorithm
Volume: 08, Issue: 01, June 2019, Page No.8-13
ISSN: 2278-2397

Algorithm- 2 : Algorithm CLOPD (X,d,k,p) Input:
// X: x1, x2……, xn be a set of training data points, d:
Distance metric (threshold),
// k: no of nearest neighbours , p: minimum no items needed to accept a cluster, iter : number of iteration
// K: number of clusters, Gr: assigned groups, Cl: cluster center point, Nout: no of outliers,
Output: Number of clusters without outliers Nout
K←0
For all x € X do [o,p]←size (X) End for
K ← 3
[Gr,Cl] ←kmeans(X, K);
For each iter < p do
for i =1 to K do for j =1 to o do
dist= Ecldistance(xi,Cli) // call Eclidean distance function Observe the cluster center point Cli
If dist < d then Add cli to cluster center point Cli Else label as outliers Nout
End if End for
End for
Compute log-likelihood for partition obtained.
For all cli consists p then
Include all cli to Cli into group Gr
End for End for
Compute mean and standard deviation for the created vector for n rows of DB Return The Real Data (DB) without outliers (Nout).

After feature selection, K-Means clustering algorithm is used for partitioning the dataset with the fixed number of clusters. Initially, number of clusters (nc) are built based on distance metrics (dist) to consider outliers (Nout) in the clustering process. Let k be the no of adjustable cluster and assign k'=k, randomly choose k' points as cluster centers C= {c1, c2... ck'). The distance 'dist' is calculated with each instance with respect to selected instance randomly. Each point of the original training dataset is to be assigned to the cluster with nearest seed. Identification of outlier in each column of the dataset is validated by undertaking kNN search upon its first object access to eliminate any false positive with threshold value given. The maximum and minimum value of clusters and the maximum distance from the centroid of the clusters are also identified. Distance of each points of the cluster from centroid is calculated based on the threshold value. If it may be greater, the points will be declared as outlier. The framework of the CLOPD method is depicted in the figure 5.1.

III. RESULTS AND DISCUSSIONS OF GBFS ON PIMA DATASET

In furtherance of exhaustive description specified in the prior chapter, different data mining classifiers like Naive Bayes (NB), Multilayer Perceptron (MLP), Sequential Minimal Optimization (SMO), RBF Network (RBFN), and random forest are implemented on the dataset PIMA using the WEKA software. The above mentioned five data mining classifier algorithms are engaged to classify dataset before and after the feature selection process. Naïve Bayes Classifier works based on Bayes theorem that provides well-built independence between the features. The PIMA dataset utilized the number of attributes selected by the GBFS algorithm and results at different phases of the algorithm are considered. With the original training dataset, a 10-fold cross validation was executed on the classifiers. The measurement are used to analysis the performance of the classifiers in numeric values like the accuracy (%), errors in performance analysis, precision rate, recall, and f–measure rate. The accuracy of every classifier with the selected optimal feature subset by the GBFS algorithm is derived. The implementations of the above declared classifiers and the outcomes are discussed in the next sections.

International Journal of Computing Algorithm
Volume: 08, Issue: 01, June 2019, Page No.8-13
ISSN: 2278-2397

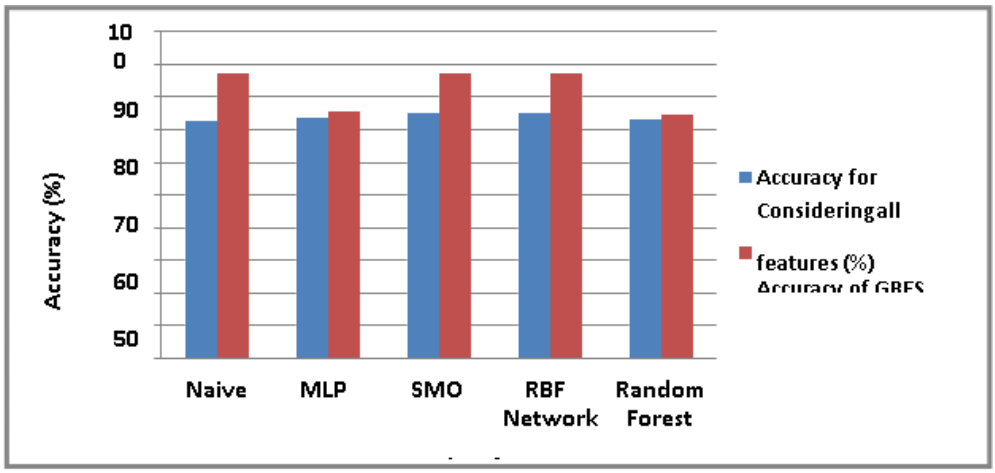

Fig-1 Accuracy comparison of GBFS algorithm

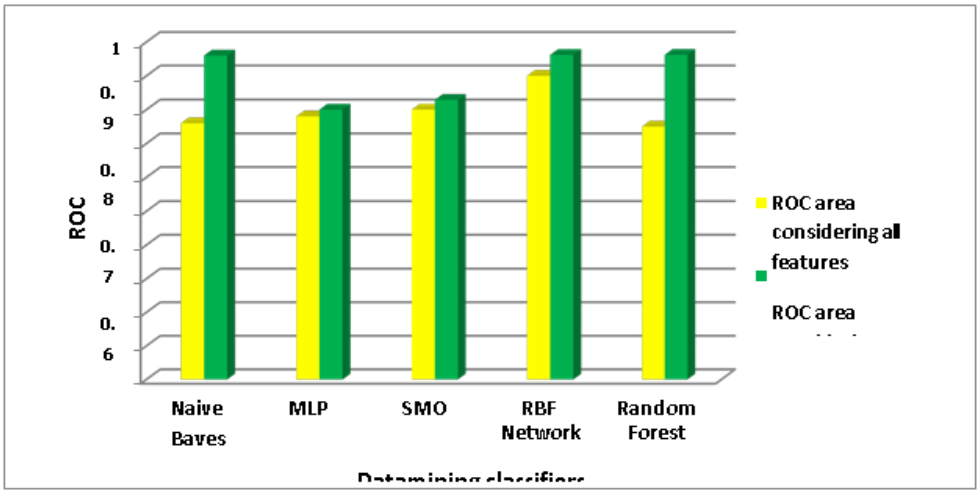

Fig-2 ROC of GBFS algorithm

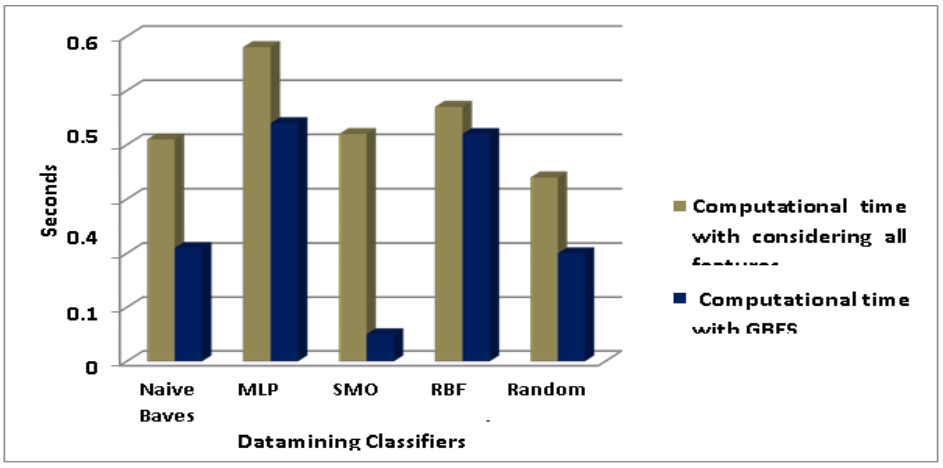

Fig-3 Comparative Analysis of GBFS with Other Feature Selection Methods

It is obvious from the table that Naïve Bayes, RBFN and Random forest used for multivariate data such as PIMA shows higher accuracy by the outcome of the GBFS methodology. The tabularization undoubtedly

International Journal of Computing Algorithm
Volume: 08, Issue: 01, June 2019, Page No.8-13
ISSN: 2278-2397

illustrates that the ROC value of the classifiers Naïve Bayes, RBFN and Random Forest are higher than other classifiers. According to the accuracy and ROC measures, it can be concluded that the classifiers of Naïve Bayes, RBFN and Random forest have improved performance on multivariate data by the utilization of GBFS methodology. The major intention of the study is to increase the prediction accuracy of cluster and the computational time of the clustering process should be reduced. Feature selection method is proposed that eliminate the irrelevant and redundant features to use different clustering techniques. Keeping significant features used for improving the speed of clustering process and quality of clustering task. The proposed GBFS algorithm is compared with existing feature selection algorithms. Correlation feature selection algorithms with various kinds of searching strategies are implemented for experimentation. The figure 4.13 indicates that the comparative analysis shows the proposed GBFS algorithm produces more accuracy than other methods.

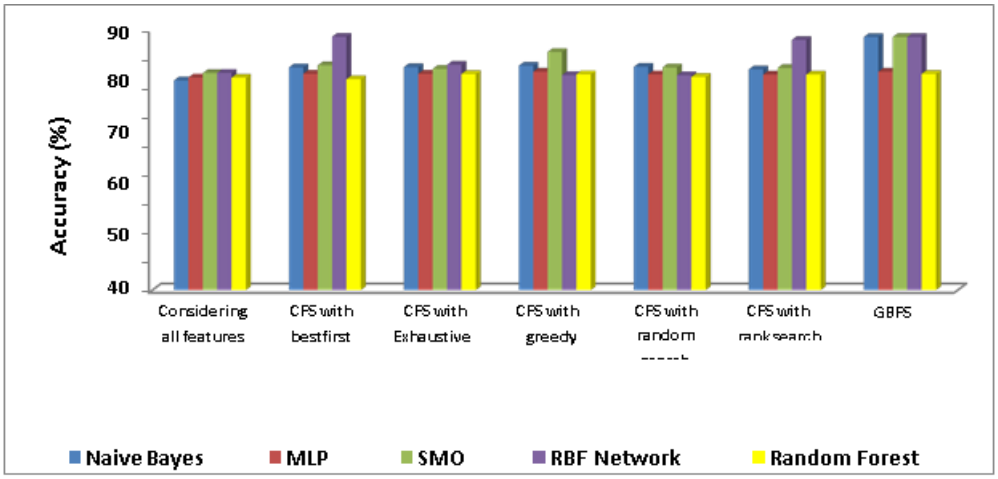

Fig-4 Comparison of Accuracy between GBFS and Other Feature Selection Methods

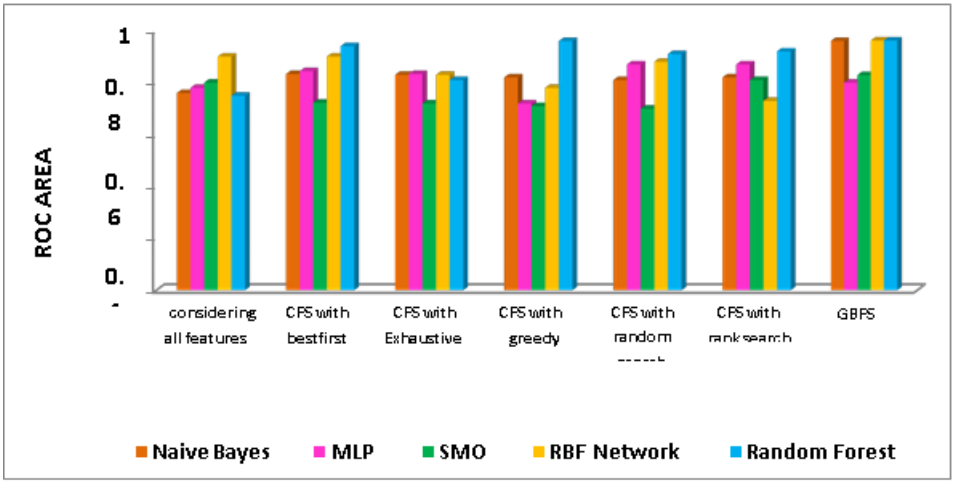

Fig-5. Comparison of ROC between GBFS and Other Feature Selection Methods

This investigation gives evidence as the proposed GBFS algorithm has gained higher values of ROC in all the five classification methods while comparing other feature selection methods. The plots of ROC area under five classifiers are clearly shown that the relevant features only produces more efficiency during classification process. This proposed GBFS method results in an extremely high true positive rate. It is concluded that the GBFS features are performed well and is suitable for cluster analysis. Therefore, these relevant features are used to consider for further clustering process. It acquires significant features for clustering process to reduce the computational time delay and comparatively more efficiency than other FS methods.

International Journal of Computing Algorithm
Volume: 08, Issue: 01, June 2019, Page No.8-13
ISSN: 2278-2397

IV. CONCLUSION

This chapter has discussed the feature selection procedure and the power of the integrated approach in the GBFS. The identification and selection of the most influenced PIMA Indian diabetes feature subset is prepared with a novel algorithm. The highlights of feature selection and relative results are carried out in this chapter. Two phases of attribute reduction in GBFS algorithm have been proposed to achieve a compact set of features for clustering process. The purpose of the approach is to choose the smallest amount of features from the obtainable feature sets and produce higher cluster compactness and homogeneity. The results of the investigational evaluation have revealed that the proposed model has reached improved classification accuracy while comparing with accuracy of all the features. The results have shown that Naive Bayes, SMO, and RBFN classifiers have reported highest accuracy of 87.5 in the PIMA dataset. The error performances of the classifiers have been proved to be minimum and the *precision*, *recall* and *F- measure* values have also been higher compared.

References

[1] Alelyani, Salem, Jiliang Tang, and Huan Liu. "Feature selection for clustering: A review." *Data Clustering: Algorithms and Applications* 29 ,pp.110-121., 2013

[2] Koufakou, Anna, and Michael Georgiopoulos. "A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes." *Data Mining and Knowledge Discovery*20, no. 2 .,pp. 259-289., 2010

[3] Anusha, M., and J. G. R. Sathiaseelan. "An improved K-means genetic algorithm for multi-objective optimization." *International Journal of Applied Engineering Research* .pp. 228-231. 2015

[4] Bermejo, Pablo, Jose A. Gámez, and Jose M. Puerta. "A GRASP algorithm for fast hybrid (filter-wrapper) feature subset selection in high-dimensional datasets." *Pattern Recognition Letters* 32, no. 5 pp. 701-711.,2011

[5] Anirudha, R. C., RemyaKannan, and NagammaPatil. "Genetic algorithm based wrapper feature selection on hybrid prediction model for analysis of high dimensional data." In *Industrial and Information Systems (ICIIS), 2014 9th International Conference on*, pp. 1- 6. IEEE, 2014.

[6] Anitha, S., and M. Mary Metilda. "A heuristic approach for observing outlying points in diabetes data set." In *Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2017 IEEE International Conference on*, pp. 199-202. IEEE, 2017.

[7] Anitha, S., and Mary Metilda An Evaluation of cluster based outlier detection strategy by feature selection technique in diabetes data set (IJPAM) Volume 119 No. 16 2018, pp.411-420

[8] Anusha, M., and J. G. R. Sathiaseelan. "An enhanced K-means genetic algorithms for optimal clustering." *Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on*. IEEE, 2014.

[9] Blake C L and Merz C J, "UCI repository of machine learning databases", 1998

[10] Shieh, Albert D., and Yeung Sam Hung. "Detecting outlier samples in microarray data." *Statistical applications in genetics and molecular biology* 8, no. 1,pp.1-24. 2009

[11] Dai, Jian-Hua, and Yuan-Xiang Li. "Heuristic genetic algorithm for minimal reduction decision system based on rough set theory." In *Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference on*, vol. 2, pp. 833- 836. IEEE, 2002.