RESEARCH ARTICLE OPEN ACCESS

Design and Implementation of an Efficient Cluster based Feature Selection System for Medical Decease Prediction using SVM Classification

D.Haveela Bala [1], Dr.M.Hanumanthappa [2]
Research Scholar [1]
Department of Computer Science
Rayalaseema University, Kurnool – 518007
Professor [2]
Department of Computer Science & Applications
Bangalore University, Jnanabharathi Campus, Bengaluru – 560056
India

ABSTRACT

The classification of medical data has become an increasingly challenging problem, due to recent advances in medical mining technology. The Clinical organizations have collected large quantities of information about patients and deceases. However in a large dataset not all features contribute to represent the disease prediction, therefore reducing and selecting a number of sufficient features may improve accuracy of the classification. The objective of this Paper is to improve the classification accuracy of the SVM based on the output from RFE feature selection method that use Clustering approach to find the optimal set of features. This approach was applied on the Two UCI datasets namely, Wisconsin Breast cancer and Pima Diabetes in python Language. Experimental results showed that using the cluster based feature selection method with SVM algorithm achieved higher classification accuracy than without feature selection SVM classifier. In our result show that features selection improve significantly the classifier performance.

Keywords:- SVM, UCI, RFE

I. INTRODUCTION

Now a day's various clinical organizations are generating huge amounts of data which are difficult to handle for further processing. The Clinical organizations have collected large quantities of information about patients, deceases and their clinical lab test results. Data mining is the search for relationships and patterns within this data that could provide useful knowledge for effective decision-making. Medical data mining is one of key issues to get useful clinical knowledge from medical databases.

Data mining is the process of extracting valid, previously unknown, and ultimately comprehensible information from large databases and using it to make crucial business decisions. The extracted information can be used to form a prediction or classification model, or to identify relations between database records [4].

Various Data Mining techniques such as classification, association rules and clustering techniques are used by clinical organization to increase their capability for making decision regarding patient health.

There are two goals when optimizing classification rules, attaining highest accuracy and selecting smallest set of features.

II. FEATURE SELECTION

Feature selection is commonly used preprocessing step of data mining that helps increase the predictive performance of a classification model [1, 2]. The main aim of feature selection is to choose a subset of features with high predictive information and eliminate irrelevant features with little or no predictive information. It is a process of selecting a subset of features that are most useful to construct a classification model. Choosing a subset of original features are important so that the feature space is optimally reduced according to a certain evaluation criterion. Feature selection is a vital issue for medical disease classification.

As far as the number of features increase, the dimensions of data increase, will affect the

classification accuracy. In fact with so many irrelevant and redundant features, most classification algorithms suffer from extensive computation time, possible decrease in model accuracy and increase of overfitting risks [7, 9]. As a result, it is necessary to perform dimensionality reduction on the original data by removing those irrelevant features.

The performance of the classifier and the cost of classification are sensitive to the choice of the features used in the building of the classifier [5]. With the reduced set of features, the time needed for learning the classification knowledge and the time required for classification is reduced. Further, by the extraction of relevant features and therefore the elimination of the irrelevant ones, the accuracy of the classifier can be increased.

Feature selection algorithms fall into three broad categories: filter model, wrapper model and embedded [5]. The filter model relies on general characteristics of the training data to select some features without involving any learning algorithm. The wrapper model requires one predetermined learning algorithm in feature selection and uses its performance to evaluate and determine which features are selected. Embedded method incorporates the feature selection process in the classifier objective function or algorithm.

The main aim of feature selection is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features. There are two goals when optimizing classification algorithms, attaining highest accuracy and selecting smallest set of features. Applying feature selection techniques in medical diagnosis has become a prerequisite for model building. The major advantages of feature selection are (i) it improves the performance of the model, (ii) it provides faster and more cost effective models and (iii) it helps gain a deeper insight into the underlying processes [6].

III. DESIGN OF PROPOSED METHODOLOGY

Figure 1 depicts the proposed system architectures of the cluster based feature selection. In this section, we present the proposed method of cluster based feature selection on SVM classification using Recursive feature elimination. We use the k-Means algorithm to find the best cluster of the features from feature ranking scores and use these results to build the SVM classification model. The objectives are to reduce the data dimensions and to increase the predictive accuracy. Our proposed method consists of the following steps.

Step 1: Read the Dataset.

Step 2: Preprocess the data

Step 3: Now partition the Data set into training and testing sets.

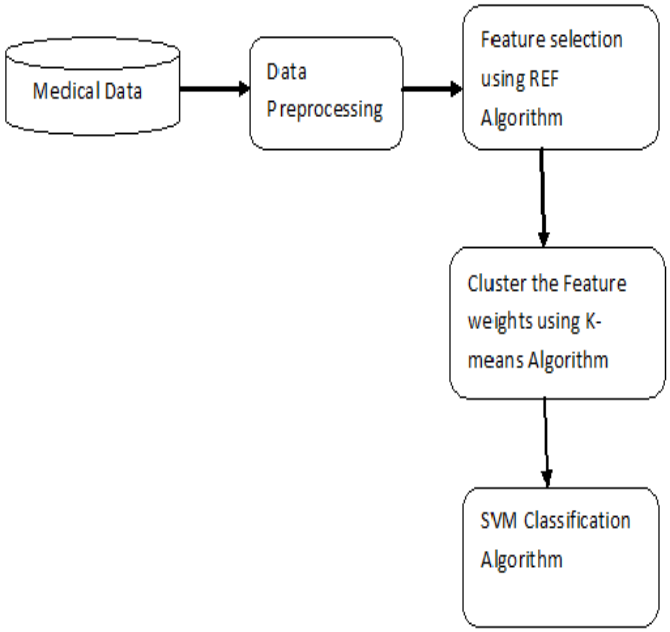

Figure 1: Cluster based feature selection architecture

Step 4: Apply RFE technique for finding the feature weights.

Step 5: Now cluster the feature weights of training data using K-means algorithm.

Step 6: Select the best cluster features.

Step 7: Train the SVM of reduced features.

Step 8: Validate the model for new classification.

3.1 SVM Recursive Feature Elimination (SVM-RFE)

Initially, RFE started with all the features. The weights w_i of a linear SVM classifier provide information about feature relevance using a decision function $D(\boldsymbol{x}) = \boldsymbol{w}\boldsymbol{x}+b$, where a bigger weight value implies higher feature relevance. For the next iteration all weights are re-evaluated and dynamically adapted, while the process continues recursively [3]. This procedure continued until all features were ranked according to the order of their removal. In

this paper a feature x_i is scored by means of w_i^2, as in the original RFE algorithm.

3.2. K-Means Algorithm

Clustering is one of the most important unsupervised learning problems widely used for exploratory data analysis. Clustering method finds a structure in a collection of unlabeled data. Hence, it separates the original dataset into smaller datasets called clusters. The K-means algorithm is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. Usually the Euclidean distance is used as the distance metric to calculate the observations' relationship. The main idea is to define k centroids, one for each cluster. The algorithm is composed of the following steps:

Step 1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

Step 2. Assign each object to the group that has the closest centroid.

Step 3. When all objects have been assigned, recalculate the positions of the K centroids.

Step 4. Repeat Steps 2 and 3 until the centroids no longer move.

Finally, each cluster could represent a different collection from the other clusters. By using this kind of clustering methods, the observations could be easily separated according to the Euclidean distance.

3.3 Support Vector Machine

Support Vector Machines (SVM) is a machine learning algorithm that is generally used for classification problems. SVM algorithm is one of the most powerful classification techniques that was successfully applied to many real world problems [2, 3]. Support Vector Machines are based on the idea of mapping data points to a high dimensional feature space where a separating hyper-plane can be found. The main logic used by SVM for data classification is to drawn optimal hyper-plane which acts as a separator between the two classes. The separator should be chosen like that it gives the maximum margin between the vectors of two classes as shown in figure 2. Due to this reason SVM is also called maximum margin classifier. The vectors near the hyper-plane are called support vectors. This mapping can be carried on by applying the kernel trick which implicitly transforms the input space into another high dimensional feature space. The hyper-plane is computed by maximizing the distance of the closest patterns, i.e., margin maximization, avoiding the problem of overfitting.

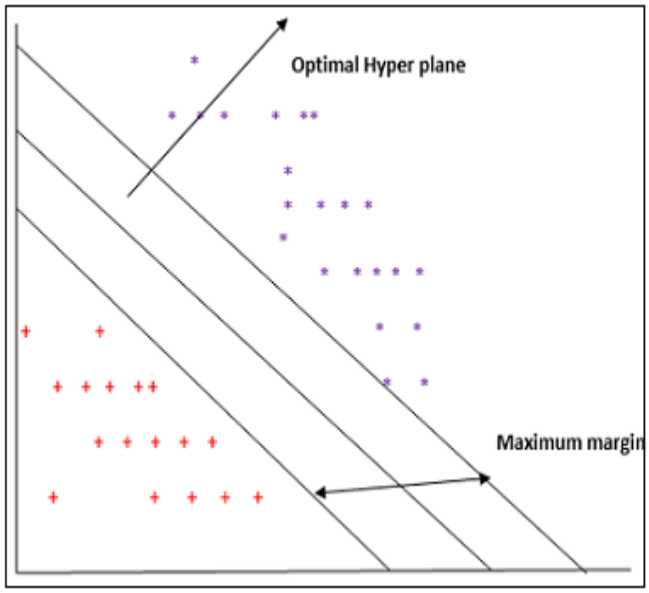

Figure 2: separating hyperplane

IV. EXPERIMENTAL RESULTS

The objective of this section is to evaluate our proposed algorithm in terms of number of selected features, and learning accuracy on selected features. The proposed cluster based feature selection methods have been experimented with data taken from the UCI Machine Learning Repository [8]. We have used the Python Language to experiment our proposed algorithms. The Python Scikit-learn is a package for data classification, regression, clustering and visualization. The dataset we used in our experiment is briefly described in Table 4.1. The data is divided in two sets. The training set is 70% and the remaining 30% are used for testing. The experiments were conducted with complete feature set and also with selected features.

Table 4.1: Summary UCI data sets.

SNO	Datasets	Features	Instances	Class
1	Wisconsin Breast cancer	11	699	2
2	Pima Diabetes	9	768	2

4.1 Evaluation of RFE algorithm

We apply the RFE method to find the weights of our Dataset results are shown in table 4.2 and 4.3.

Table 4.2: Wisconsin Breast cancer Data feature weights

Features	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
Weights	0.034	0.26	0.22	0.025	0.13	0.17	0.098	0.073	0.06

Table 4.3: Pima Diabetes Data feature weights

Features	preg	plas	pres	skin	ins	mass	pedi	age
Weights	0.082	0.26	0.086	0.069	0.074	0.17	0.125	0.146

Next we apply the k-means algorithm for find clusters according to feature weights of RFE algorithm are shown in table 4.4 and 4.5

Table 4.4: cluster of Wisconsin Breast cancer Data

	Cluster-1				**Cluster-2**				
Features	Uniformity of Cell Size	Uniformity of Cell Shape	Single Epithelial Cell Size	Bare Nuclei	Clump Thickness	Marginal Adhesion	Bland Chromatin	Normal Nucleoli	Mitoses
Weights	0.26	0.22	0.13	0.17	0.034	0.025	0.098	0.073	0.06

Table 4.5: cluster of Pima Diabetes Data

	Cluster-1				**Cluster-2**			
Features	plas	mass	pedi	age	preg	pres	skin	ins
Weights	0.26	0.17	0.125	0.146	0.082	0.086	0.069	0.074

Table 4.6 shows comparative results of classification accuracy and the same shown in bar graph in figure 2. It can be seen that the SVM algorithm of all features of accuracy on Breast cancer (95%) and Pima Diabetes (75%) data sets. The proposed cluster based feature selection on SVM algorithm can improve the performance of accuracy on Breast cancer (98%) and Pima Diabetes (96%) data sets when compared to data set with no feature selection method.

Table 4.6 Result of SVM

Datasets	Accuracy of all features	Accuracy of selected features
Wisconsin Breast cancer	95%	98%
Pima Diabetes	75%	96%

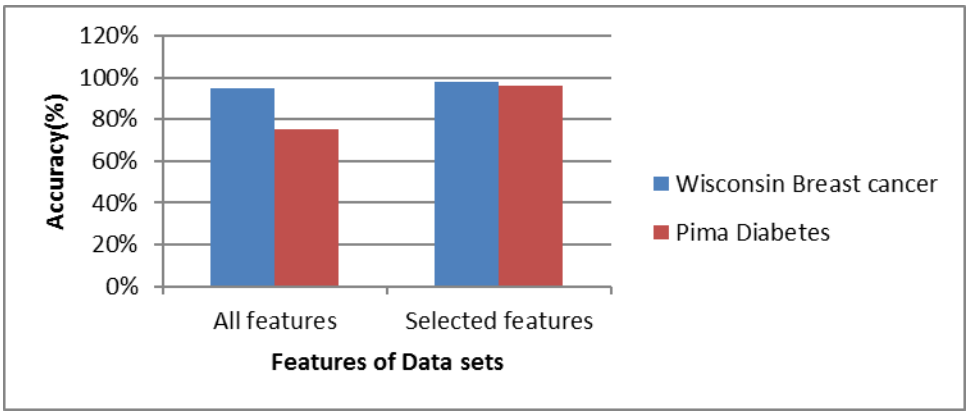

Figure 2: Accuracies of all features and cluster based features

V. CONCLUSION

This research aims at studying a method to clustering the feature ranking on SVM classification using cluster based feature selection. The problem of learning efficient model from data with high dimensionality can cause trouble to most algorithms. Thus, we propose to use the cluster based feature selection algorithm in order to increase accuracy and reduce learning problem due to dimensionality. We present clustering method using the k-Means algorithm to cluster the feature ranking scores for choosing an optimal set from feature ranking score.

From experimental results, it has been revealed that the proposed cluster based feature selection method can increase the accuracy of data classification, and can reduce high dimensional data problem by obtaining a small set of features.

REFERENCES

[1] A.L.Blum and P.Langley, "Selection of Relevant Features and Examples in Machine Learning, Artificial Intelligence", Vol.97, PP: 245-271, 1997.

[2] 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.

[3] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," Mach. Learn., vol. 46, no. 1-3, pp. 389–422, 2002.

[4] J.Han and M.Kamber,"Data Mining concepts and Techniques", the Morgan Kaufmann series in Data Management Systems, 2nd ed.San Mateo, CA; Morgan Kaufmann, 2006.

[5] R. Kohavi, G. John, Wrapper for feature subset selection, Artif. Intel., Vol. 97 , pp. 234–273, Dec.1997.

[6] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007 Oct 1; 23(19):2507±17. doi: 10.1093/bioinformatics/btm344 PMID: 17720704

[7] T. Howley, M.G. Madden, M. L. O'connell, A.G. Ryder, "The effect of principal component analysis on machine learning accuracy with highdimensional spectral data. ", Knowl.-Based Syst., vol. 19, num. 5, 2006, pp. 363-370.

[8] UCI machine learning repository. http://archive.ics.uci.edu/ml/.

[9] Y. liu, M. Schumann, "Data mining feature selection for credit scoring models", Journal of the Operational Research Society, vol. 56, num. 9, 2005, pp. 1099–1108.