4th International Conference on Science, Technology and Management
India International Centre, New Delhi (ICSTM-16)
15th May 2016, www.conferenceworld.in ISBN: 978-81-932074-8-2

FEATURE EXTRACTION TECHNIQUES USING SUPPORT VECTOR MACHINES IN DISEASE PREDICTION

Sandeep Kaur[1], Dr. Sheetal Kalra[2]

[1,2] *Computer Science Department,*

Guru Nanak Dev University RC, Jalandhar(India)

ABSTRACT

Data mining process is becoming important in healthcare industry due to very large volume of data produced and collected by them on daily basis. Support Vector Machine is the most commonly used classification algorithm for disease prediction in healthcare industry. It is widely used to predict the disease like diabetes, breast cancer, lung cancer, heart disease etc. It is advantageous to reduce the number of input features to Support Vector Machine in order to get efficient results. To reduce feature set, the aim is to select only the useful features from the entire set of features. There are many methods available for feature extraction. In this paper, various feature selection or extraction methods like F-score, Genetic Algorithm, K-means, ReliefF and SVM-RFE are discussed for disease prediction using Support Vector Machines. This paper also gives the comparison of accuracy and efficiency achieved by these feature extraction techniques for predicting various diseases.

Keywords: Data mining, F-score, Genetic Algorithm, K-means, ReliefF, Support Vector Machines.

I. INTRODUCTION

Data mining is the process of representing the useful and meaningful data from very large volumes of data. It is also referred to as 'knowledge mining' i.e. to obtain knowledge from the data. Data mining gives solution to the problems by analyzing the data that is stored in the database [1]. It is highly helpful in healthcare industry to predict various diseases. Data mining process is becoming important because healthcare industry is highly rich with information. On daily basis, a very large volume of data is produced and collected by them. By extracting useful information only, work efficiency is improved and the quality of decision making process is enhanced [2]. Classification plays a vital role in the process of data mining. Classification divides the data in a collection, into categories or classes. The main aim of classification methods is the accurate prediction of the target class or category for each item in the data. The classification algorithm analyzes relation between the data and the target class or category. Different classification algorithms have different techniques to find this relationship [3]. For predicting the results, the algorithm uses a training set which contains a set of attributes and the target outcome or class. The classification algorithm is then given a dataset that is not seen before; this dataset is called

prediction set. The algorithm then analyzes this input dataset and produces a resultant prediction. The accuracy of prediction defines that how good this algorithm is [4].

Support Vector Machine (SVM) is the most commonly used classification algorithm for disease prediction. It is widely used to predict diabetes, breast cancer, lung cancer, heart disease etc. It is a supervised learning technique that is used for discovering patterns for classification of data [5]. SVMs were first introduced by Vapnik in 1960s for classification of data. The two elements used for the implementation are the mathematical programming and the kernel functions. The kernel function allows it to search for a variety of the hypothesis spaces. In SVM, classification is performed by drawing hyperplanes. In two class classification, this hyperplane is equidistant from both the classes. The data instances which are used to define this hyperplane are known as support vector. A margin is defined in SVM which is the distance between hyperplane and the nearest support vector. For good separation by this hyperplane, the distance of margin should be as large as possible because large distance gives less error. If the margin is close then it is more sensitive to noise [6].

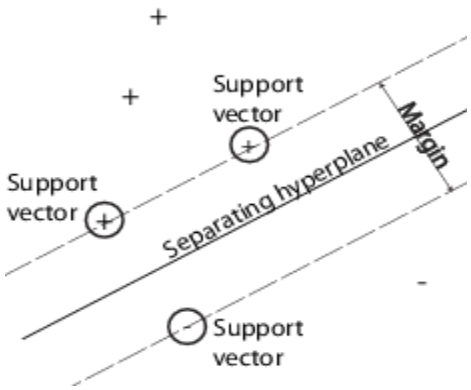

Fig 1: SVM hyperplane

The equation to define the hyperplane and the margin are $w^t\ x + b = 0$ and $w^t\ x + b = \pm 1$ respectively. Here, w is defined as a weight vector and b as bias.

For better results of SVM, the features that are given as an input to SVM are needed to be reduced. The reduced feature set helps to improve the efficiency of the results produced by the algorithm. To reduce features set, only the useful features are selected from the entire set of features. In feature selection, there is a set of features and a method is used to select a subset of features that can perform best under the classification system. The term 'feature selection' refers to the algorithms that gives a subset of feature set which are given as an input to the algorithm [7]. The main three approaches of feature selection are filter, wrapper, and embedded. In Filter methods, high ranked features selected on the basis of a statistical score. In wrapper and embedded methods, the design of a classifier is considered to select the subset of features [8]. The various feature selection or extraction methods like F-score, GA, K-means, ReliefF and SVM-RFE are discussed in next subsections. These all techniques are helpful to efficiently and easily extract the features for Support Vector Machines.

1.1 F-score

F-score is a simple and generally effective method to select the useful features from the dataset. It is the technique that measures the differentiation of two real number datasets. F-score for each feature is calculated

according to the positive and negative instances of that feature. The features having high F-score value are selected and then the SVM is applied on them for prediction. The equation for calculating F-score is:

$$F_i = \frac{(\bar{x}_i^{+} - \bar{x}_i)^2 + (\bar{x}_i^{-} - \bar{x}_i)^2}{\frac{1}{n_{\pm} - 1} \sum (x_{k,i}^{\pm} - \bar{x}_i^{\pm})^2}$$

Where k is the positive or negative instance of the i^{th} feature. The numerator defines the differentiation of the positive and the negative sets. Firstly, calculate the F-score for each feature and then set a threshold value which has lowest validation error. Eliminate the features having F-score value less than the selected threshold value. The only disadvantage of using F-score is it does not provide any mutual information among the features [9].

1.2 Genetic Algorithm

Genetic Algorithm is a searching technique that basically works on the principle of genetics to generate the useful solutions to search problems. It generates solutions using inheritance, selection, crossover and mutation. In selection, a string is selected on the basis of fitness function to generate a new generation; In crossover, the good strings are combined to generate offspring; In mutation, string is altered to conserve genetic diversity from one generation to the next. The population is repeatedly modified in each generation until the specific termination criterion is satisfied. In selection process, the fitness function is dependent on the problem and represents the solution either in 0 or 1. The important termination conditions are satisfaction of minimum criteria by the solution, budget reached, required number of generations reached, highest ranking fitness of the solution reached etc. In this, the better solution is chosen by comparing it with other solutions [10, 11].

1.3 K-means

K-means is an unsupervised learning algorithm to classify the objects on the basis of features into k groups. The very first step in k-means is to identify the number of clusters or groups i.e. *k*. Then, the *k* centroids are randomly selected for each cluster. The objects are assigned to centroid by calculating the distance of every point from every centroid. The object is assigned to the minimum distant cluster. When all the objects have assigned the clusters then it again calculate new centroid by taking mean of objects in the cluster. These steps are repeated until none of the object in the cluster changes the centroid based on its new distance from every new centroid point [12]. It is the most simplest, effective and easily understandable method for feature selection. Its efficiency depends on the selection of number of clusters and randomly selected centroids. It finds a partition where the objects within each cluster are close to each other and far from the objects in other clusters. Its main motive is find the clusters in such a way that there is always less intercluster similarity and more intracluster similarity.

1.4 ReliefF Algorithm

It is one of the successful algorithms for feature extraction in machine learning. It extracts the features at random by computing the nearest neighbours [13]. The features are extracted in a way that there is more interclass difference and less intraclass difference. After selecting random data, it finds k nearest hits and misses. 'Hits' are the features having same class label and 'misses' are the features having different class label.

It then adjust the features according to the difference of selected data's feature values and the nearest neighbors. The instances with their class label are given as a input to this algorithm. Initially, the weight say '$w(f)$' associated with each feature is set to be zero. Then in iterations, it randomly selects the instance and finds its nearest hits and misses. Then, the value of $w(f)$ is updated, as it increases if value of randomly selected instance is different from the nearest miss and decreases if its value is different from nearest hit as identified from the following equation.

$$w(f) = w(f) - (\text{selected instance} - \text{nearest hit})^2 + (\text{selected instance} - \text{nearest miss})^2$$

According to the values of $w(f)$, the ranking of each feature is identified and then features are selected by eliminating those features which have smallest weight values [14,15].

1.5 SVM-RFE

SVM—RFE stands for Support Vector Machine- Recursive Feature Elimination. It is an embedded approach that recursively removes unimportant features rather than using the weights for ranking criterion as in Relief F. It helps to provide better performance by selecting best features subset. It takes training instances and their class labels as an input to the algorithm and uses ranking criterion based on the weight vector of SVM. From the weight vector of SVM, the ranking of each feature is identified and then features are selected by eliminating those features which have lowest ranking. In bioinformatics, it is a powerful feature selection algorithm to avoid overfitting in case of high number of features. But this algorithm can only be used to linear kernel SVM, because in case of non linear kernel it is quiet difficult to find the weight vector [15].

II. RELATED WORK

Khyati K. Gandhi, Prof. Nilesh B. Prajapati in 2014 performed[8] feature selection techniques on diabetes data set (Pima Indian diabetic database) from UCI repository. F-score, ReliefF and Genetic Algorithm are used for feature selection from the diabetes dataset and then the classification is performed by using Support Vector Machine classifier. It has been analyzed that the performance of SVM is better enhanced by using F-score technique on diabetes dataset. The accuracy achieved by F-score is more than the other methods. The accuracy of Genetic Algorithm is analyzed by using Support Vector Machine as well as by Artificial Neural Networks. The result shown that the accuracy achieved is more in case of SVM.

Xiaobo Li *et al* in 2011 presented [16] a comparison of seven different feature selection techniques on multiclass cancer dataset. The seven feature selection methods are Correlation based, Chi-Squared, Gain Ratio, Information Gain, ReliefF, SVM-RFE and Symmetrical Uncertainty. The experimental results show that the feature selection by using SVM-RFE gives better performance than other six methods. The feature selection on multiclass cancer is critical, but it is possible to achieve better accuracy on the dataset by using proper feature selection and classification methods.

B Zheng *et al* in 2013 proposed [17] a model that is a hybrid of K-Means and Support Vector Machine. The model is implemented on breast cancer dataset to diagnose cancer based on the extracted features of tumor. K-means is used for finding the hidden pattern of tumor and SVM for classification of features. There are two types of tumors: malignant (are cancerous) and benign (can't be cancerous, can be removed). The classifier separates these two types of tumors. The k-means is used for clustering the patterns of the similar tumor based

on the features of malignant and benign tumors. The membership function is used to measure the similarity of the data point and the tumor. The results show that the K-means and SVM hybrid model reduces the time required for prediction with higher rate of accuracy.

G.Ravi Kumar *et al* in 2014 proposed [9] a hybrid of Genetic Algorithm and Support Vector Machine for large medical datasets. The main aim of the proposed model is to select the features automatically for SVM classifier. Genetic Algorithm selects the best set of features by eliminating insignificant features based on the fitness function. The proposed model is implemented on five different disease datasets from UCI repository. The performance of proposed model is compared with the performance of SVM classifier alone which has no feature extraction method. The proposed work improved the accuracy and performance of classification as compared to the SVM approach.

Cheng-Lung Huang *et al* in 2008 proposed [18] a hybrid model of F-Score and Support Vector Machine for breast cancer dataset. The main aim of this paper is to get the bioinformatics about cancer tumor and DNA viruses. The features are selected by using F-score method and the Support Vector Machine parameters are sharpened by grid search. This paper gives five DNA viruses that are HSV-1, CMV, EBV, HHV-8 and HPV which affect a breast tumor being diagnosed by SVM. To find the correlation of DNA virus with the tumor and to achieve better accuracy, F-Score is adapted for feature extraction in this case. The results show that the model helped to achieve good accuracy in predicting the breast cancer.

III. COMPARISON OF VARIOUS FEATURE EXTRACTION METHODS

In this study, the accuracy of F-Score, Genetic Algorithm, K-means, ReliefF and SVM-RFE feature extraction techniques is compared on breast cancer WDBC (Wisconsin Diagnostic Breast Cancer) data set from UCI repository. The accuracies achieved by these methods is shown in the TABLE 1.

In the United States, breast cancer is becoming the main reason of death in women. Every 1 out of seven women is diagnosed with cancer. The most recommended way to get rid of breast cancer deaths is to detect it in its earlier stage. Breast cancer is detected on the basis of the type of tumor in breast. There are two types of tumors: malignant (are cancerous) and benign (can't be cancerous, can be removed) [19]. The features of tumor are selected by various feature extraction techniques in order to identify the chances of breast cancer.

Table 1: Comparison of feature extraction techniques.

Method	Description	Classifier	Accuracy
F-Score	F-score for each feature is calculated and then features having high F-score value are selected for SVM classifier.	SVM	86%
Genetic Algorithm	It works on the principle of genetics to generate the useful solutions through inheritance, selection, crossover and mutation.	SVM	94%

K-means	It is an unsupervised learning algorithm to classify the objects on the basis of features into k groups. It finds a partition where the objects within each cluster are close to each other and far from the objects in other clusters	SVM	96%
ReliefF	It randomly selects the instance and finds its nearest hits and misses. Then, the value of weights assigned to each feature is updated. According to the values of weights, the ranking of each feature is identified	SVM	87%
SVM-RFE	Support Vector Machine- Recursive Feature Elimination. It is an embedded approach that recursively removes unimportant features rather than using the weights for ranking criterion as in Relief F.	SVM	97%

As the TABLE shows the highest accuracy of feature extraction for Support Vector Machine is achieved by SVM-RFE that is 97%. K-means being the simplest method to extract features also helps to achieve good accuracy of 96%. So, it is beneficial to use any of these two methods in order to extract features for breast cancer diagnosis.

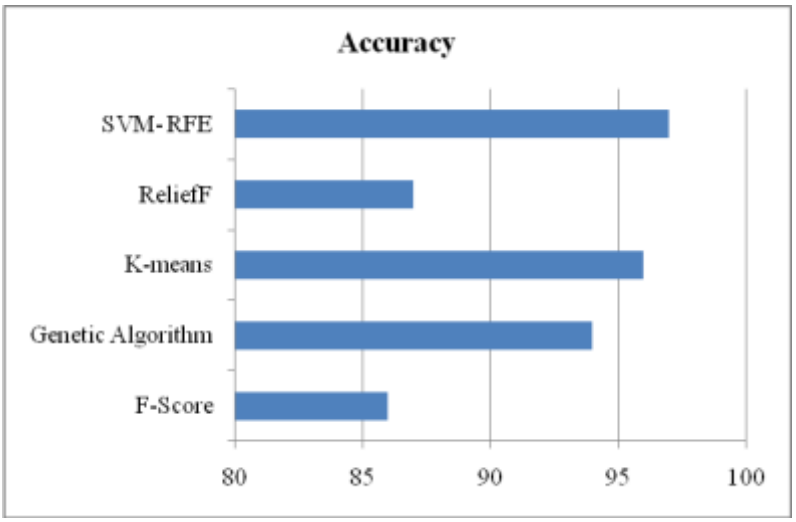

Fig 2: Accuracy Graph

IV. CONCLUSION

Support Vector Machine (SVM) is the most commonly used classification algorithm for disease prediction. For better results of SVM, the features that are given as an input to SVM need to be reduced. To reduce features set, only the useful features are selected from the entire set of features. This paper gives the comparison of various feature extraction techniques like F-Score, Genetic Algorithm, K-means, ReliefF and SVM-RFE in terms of accuracy achieved by them in order to predict the tumor in breast diagnosis. SVM-RFE helps to achieve the highest accuracy of 97% and K-means being the simplest method helps to achieve the accuracy of 96%. So, it is

beneficial to use any of these two methods in order to extract features for breast cancer diagnosis. Our future work is to study the various enhancements of K-means, and then use the best enhancement, in terms of performance, as feature extraction method for breast cancer dataset.

REFERENCES

[1] Jain, N., & Srivastava, V. (2013). Data Mining techniques: A survey paper.*IJRET: International Journal of Research in Engineering and Technology*,*2*(11), 2319-1163.

[2] Milovic, B., & Milovic, M. (2012). Prediction and decision making in Health Care using Data Mining. *Kuwait Chapter of the Arabian Journal of Business and Management Review*, *1*(12), 126.

[3] https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm

[4] Voznika, F., & Viana, L. Data Mining Classification.

[5] http://www.enggjournals.com/ijet/docs/IJET13-05-03-292.pdf

[6] Burbidge, R., & Buxton, B. (2001). An introduction to support vector machines for data mining. *Keynote papers, young OR12*, 3-15.

[7] Jain, A., & Zongker, D. (1997). Feature selection: Evaluation, application, and small sample performance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *19*(2), 153-158.

[8] Gandhi, K. K., & Prajapati, N. B. (2014, August). Study of Diabetes Prediction using Feature Selection and Classification. In *International Journal of Engineering Research and Technology* (Vol. 3, No. 2 (February-2014)). ESRSA Publications.

[9] Chen, Yi-Wei, and Chih-Jen Lin. "Combining SVMs with various feature selection strategies." *Feature extraction*. Springer Berlin Heidelberg, 2006. 315-324.

[10] RKumar, G., Ramachandra, G. A., & Nagamani, K. (2014). An Efficient Feature Selection System to Integrating SVM with Genetic Algorithm for Large Medical Datasets. *International Journal*, *4*(2).

[11] Li, X. Gene selection for cancer classification using the combination of SVM-RFE and GA.

[12] Teknomo, K. (2006). K-means clustering tutorial. *Medicine*, *100*(4), 3.

[13] Wang, Y., & Makedon, F. (2004, August). Application of Relief-F feature filtering algorithm to selecting informative genes for cancer classification using microarray data. In *Computational Systems Bioinformatics Conference, 2004. CSB 2004. Proceedings. 2004 IEEE* (pp. 497-498). IEEE.

[14] https://en.wikipedia.org/wiki/Relief_(feature_selection)

[15] Cho, B. H., Yu, H., Kim, K. W., Kim, T. H., Kim, I. Y., & Kim, S. I. (2008). Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods. *Artificial intelligence in medicine*, *42*(1), 37-53.

[16] Li, X., Peng, S., Zhan, X., Zhang, J., & Xu, Y. (2011, October). Comparison of feature selection methods for multiclass cancer classification based on microarray data. In *Biomedical Engineering and Informatics (BMEI), 2011 4th International Conference on* (Vol. 3, pp. 1692-1696). IEEE.

[17] Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Systems with Applications*, *41*(4), 1476-1482.

[18] Huang, C. L., Liao, H. C., & Chen, M. C. (2008). Prediction model building and feature selection with support vector machines in breast cancer diagnosis. *Expert Systems with Applications*, *34*(1), 578-587.

[19] Gupta, S., Kumar, D., & Sharma, A. (2011). Data mining classification techniques applied for breast cancer diagnosis and prognosis. *Indian Journal of Computer Science and Engineering (IJCSE)*, *2*(2), 188-195.