An Improved Feature Selection and Classification of Gene Expression Profile using SVM

Kavitha KR, Aiswarya Rajan KV, and Anjali Pillai
Department of Computer Science and Applications
Amrita Vishwa Vidyapeetham
Amritapuri
India
kavithakr@am.amrita.edu,aiswaryapoduval009@gmail.com,pillai.anjali18@gmail.com

Abstract: **Support Vector Machine (SVM) is the most widely used classifier for performing the classification of a massive dataset. This research paper aims to improve the feature selection and classify the gene expression data by using the SVM classifier. And also aim to decrease the computational time of the SVM-RFE (support vector machine recursive feature elimination) algorithm by identifying more than one redundant genes and removing them in every iteration. Most of the gene expression profile contains an enormous number of features with few numbers of samples, to reduce the number of features before applying to the classifier for performing the classification; a feature selection algorithm is needed. The most effective algorithm used to perform the feature selection of microarray is, SVM-RFE. On every iteration, it generates the rank of the features and removes the very least ranked feature, which is the most irrelevant. Since the modified algorithm is used to remove more than one redundant features in every iteration. It will help to reduce the computational time and increase the accuracy prediction.**

Keywords—gene expression, embedded feature selection, svm, svm-rfe, redundancy, correlation

I. INTRODUCTION

In the DNA microarray technology, gene expression profiling is the main tool used for measure the information about mRNA of thousands of genes at a time over disease or treatment and it shows the gene expression profile has great potential for predicting the distinct type of cancers [1]. It helps the scientists to identify whether the expression level of thousands of gene shows changes towards a disease or a treatment. The gene expression profile contains thousands of features with few numbers of samples. When a high dimensional dataset is used for classifier generation it takes more computational time in classification and reduces the accuracy of the classifier. So we need an efficient feature selection (FS) algorithm to reduce the dimension of the dataset.

The SVM-RFE is the most efficient FS algorithm proposed by Guyon [2]. This algorithm generates the ranks of the feature and removes one irrelevant least rank feature in a single iteration. In this case, the computational time is increased. Since this research paper aims to decrease the computational time of the SVM-RFE algorithm by identifying more than irrelevant features and removing them in each iteration. And also aims to improve the feature selection and classify the gene expression data by using SVM. In fast correlation-based filter, it removes correlated genes from the dataset and also improves the performance of SVM-RFE [3].

The feature selection is a dimensionality reduction technique, which removes the irrelevant data and improves the accuracy prediction ability of the classifier by reducing the size of the dataset. LDA (Linear discriminative analysis) is also a commonly used technique for performing the dimensionality reduction [4]. Feature selection means selection of most relevant feature subsets from the original feature sets [5][6]. There are different categories of feature selection methods, they are wrapper, filter, embedded, hybrid and ensemble methods. Based on these five FS methods, the embedded method is the most efficient one. Because rather than other methods, it has only one phase that incorporates the feature selection technique of the learning algorithm and uses its characteristics for feature evaluation [7]. And also the method is doesn't use the classifier repeatedly and features subset evaluation. Based on this reason it takes a short time at the time of computational. So in our research paper, we took the embedded method as the efficient feature selection method for performing the feature selection. The widely used embedded method is SVM-RFE and it helps to eliminate the features recursively. Below Fig.1 shows the embedded feature and classification method employed in this research. The embedded FS method selects the feature based on the output of the classification.

The proposed method uses an efficient classifier to perform the classification, which is SVM. The SVM is the most powerful supervised learning technique that constructs a model that is used to predict where the new samples fall either in positive class or negative. The SVM builds a set of hyperplanes in a high dimensional space, and a good separation of the hyperplanes is achieved based on the largest

978-1-7281-0283-2/19/$31.00 ©2019 IEEE

distance to the nearest training data point of any class. The prediction is only done on the basis of training data.

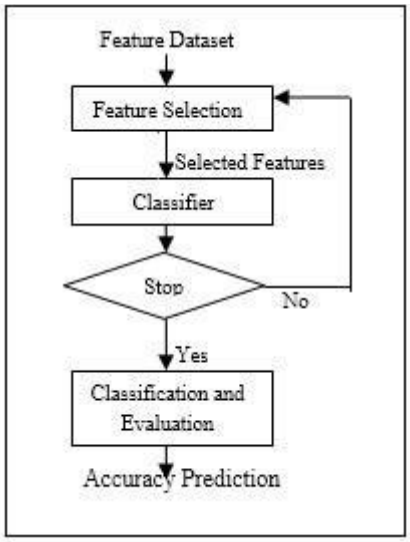

Fig. 1. Embedded Feature Selection Method

II. RELATEDWORKS

Jiapeng Yin at el [8] describes how to improve the performance of SVM-RFE. Mainly the RFE techniques are used to remove the irrelevant features. But in some cases, it can't be dealing with the redundant features. So this research paper introduces a new FS method to overcome this drawback of SVM-RFE, which is correlation among features. By using this method the redundant features are removed. This proposed method is performing the testing, based on the pancreatic cancer dataset. This dataset contains more than fifty thousands of features. By performing the classification in this massive amount of genes, at a time the SVM-RFE is removing one feature. In this case, it takes more computational time and decreases the accurate prediction. The main focus of this research is to improve the classification performance of pancreatic cancer dataset. According to this, two strategic steps are used to removing the features. First, in each iteration the bottom-ranked features (N) will remove; remaining features will take the 2's power, then again remove the half of the features. Next, the remaining features satisfy N greater than 1000, then every time 100 features are removed. But in the case of N less than 1000, at a time only one feature is removed. Based on these experimental results the proposed method has an excellent performance on the baseline of SVM-RFE.

In this paper [9] Ying Zhang is told about multiple SVM technologies with gene expression data. This research introduces a new method to find relevant features from gene expression data named as SVM-RFE based on parameter optimization (SVM-RFE-PO). For finding optimal parameters in the feature subset selection stage of SVM-RFE-PO uses other algorithms like Grid Search with SVM-RFE, Particle Swarm Optimization on SVM-RFE and genetic algorithm. Then the selected feature subset is used to train the SVM classifier. The effect of SVM-RFE-PO also compares with feature selection methods like random forest FS (RFFS), RFFS with grid search and minimal redundancy maximal relevance. The methodologies are used; the RFE algorithm constructs a ranking based on their weight vector generated by the SVM classifier. In each iteration, it removes the signature features with the lowest coefficient rank, and finally obtains all signature features that sorted in decreasing order. In machine learning, the random forest is an effective prediction tool [10]. At the training time it generates many more decision trees. According to these trees output, the result of classification is generated. Each time RFFS removes the least required feature from the feature set. After selecting all relevant features the classification accuracycalculated.

In this paper [11] Zifa Li and Weibo Xie describe the classification and feature selection. This paper mainly focuses on one of the best FS methods, which is SVM-RFE. The major drawback of SVM-RFE is large time-consuming. So this paper introduces the linear support vector machines to overcome this drawback of SVM-RFE. This is used to rank the features by training the SVM classification model. The major methodology is used, Large-scale linear support vector machines (LLSVM). In this method, the SVM is the best classifier used for performing the classification of microarray data. This paper introduces a large scale linear SVM to replace the basic SVM. For performing the large-scale classification tasks, here the LLSVM is used. This is also performing the text classification. By applying the LLSVM to microarray data, this is similar to SVM for performing the feature weighting and exceptionally fast at the time ofclassification.

In Varsha J at [12] describes the classifier for the breast cancer genes. The breast cancer dataset that contains 47,294 genes and 128 samples. This massive amount of dataset is loaded to the classifier there is a chance to increase the computational time while performing the classification. So there is a need to reduce the computational time. To overcome this drawback the research introduces the modified algorithm to perform the classification on this dataset, which is SVM-RMFE. Before applying the SVM-RFE a correlation based SVM-RMFE is used here. The SVM-RMFE is used to rank the features and remove multiple irrelevant feature form the dataset. It helps to reduce the time and increase the prediction ability of the classifier. The major methodologies are used in this research are virtual gene and SVM-MRFE. In the virtual gene algorithm, the genes are taken to a pairwise and it helps to identify the class label. Based on these methodologies the research has got the most accurate result while performing the classification of breast cancer dataset.

III. METHODOLOGY

The main aim of this research paper is to improve the feature selection and classification for gene expression profile that lowers the computational time of the classifier when a new gene expression arrives and overall accuracy of the classifier should increase. SVM-RFE is the widely used embedded method [13] and the most effective algorithm used

in the case of feature selection for gene expression dataset. SVM-RFE calculates the weight of features in each iteration, and rank them based on the weight and removes one lowest rank feature. So it is time-consuming and it removes only the redundancy features which in turn can decrease the prediction accuracy. To overcome both of these drawbacks our proposed method introduces an improved SVM-RFE.

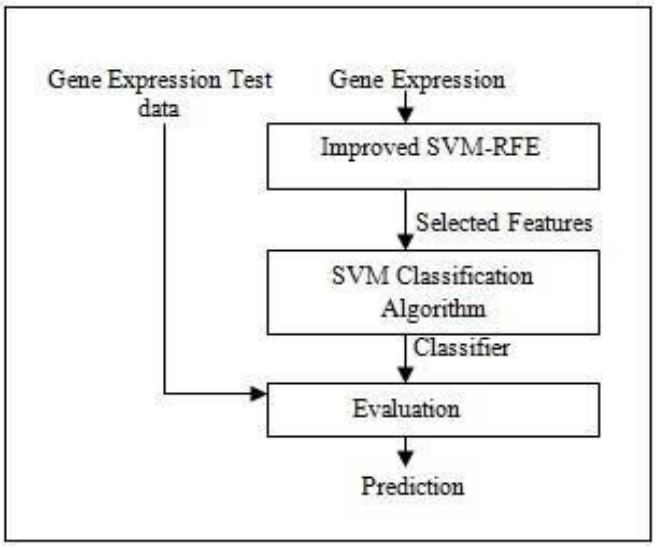

Fig. 2. System Architecture

This proposed method uses the embedded method for performing the feature selection. There are five FS methods. Like Filter, Wrapper, Ensemble, Hybrid and Embedded methods. Embedded based FS method involves the feature selection module based on the classifier. The output of the classifier is used as a criterion for ranking the features. Among all these methods the embedded is the most effective method. Also, this proposed method has an additional stage rather than other methods which is stopping criteria, the criteria is when a search is completed. Until the search is completed, the best feature subset is retrieved and perform the evaluation and classification process.

ISVM-RFE is an improved support vector machine recursive feature elimination algorithm that initially starts with a full set of features dataset. In each iteration a set of redundant features are removes based on the threshold value. This process is continued until all the features are ranked and a subset of features is selected which increase the accuracy. On improved SVM-RFE, the whole dataset is trained using SVM and calculate the weight of each feature. The weight of the features sorted in ascending order and the index of features are sorted according to their weight. The features removed first from the array(feature set) based on weight gets rank first. This procedure repeats until when the feature set becomes empty and gets rank for all features. In this way, improved SVM-RFE generates rank. By using improved SVM-RFE, it reduces the lowest ranked features, which is the most irrelevant. This is helpful to avoid computational time and increase the accuracy prediction ability of the classifier. Our proposed solution is represented in Fig.2.

ISVM-RFE is improved to remove a set of a redundant gene from the dataset based on the threshold. Rather than simply removing some *x%* of low ranked features, in our proposed method where this removal is done efficiently and effectively. The threshold is find by repeatedly finding a weight starting from the mid-weight of the sorted weight of genes and then recursively dividing the lowest weighted genes until a mid weight is reached such that the difference of the mid-weight with its lowest weight of the genes is less compared to the weight difference between the mid and the highest weighted gene. Then remove all the genes below the threshold in each iteration. The algorithm to find the threshold is given below:

Algorithm 1: feature_remove_threshold

Input: = $\{w_1, w_2, w_3, \ldots, w_n\}$ sorted in increasing order where n is the n^{th}gene.

Output: the threshold of the features to remove

1. Find the midweight
   $W_{mid} = (W_n - W_0)/2$
2. Calculate left difference from the midweight
   $left\_diff = W_{mid} - W_0$
3. **Calculate the right difference from the midweight**
   $right\_diff = W_n - W_{mid}$
4. **Repeat until right_diff<left_diff**
   a. Update the W = $\{w_0, w_1, \ldots, w_{mid}\}$
   b. UpdateWn=Wmid
   c. Find the newW_{mid}
      $W_{mid} = W_0 - W_n$
   d. Calculate left difference from the mid weight
      $left\_diff = W_{mid} - W_0$
   e. Calculate the right difference from the mid weight
      $right\_diff = W_n - W_{mid}$
5. ReturnW_{mid}

The threshold selection method is called in SVM-RFE. In this the features are ranked based on weight, and **feature_remove_threshold (w)**is called based on the current weight and a threshold is returned and all the features below the threshold are removed in each iteration. Hence decreases the computation time in ISVM-RFE.

Algorithm 2: ISVM-RFE

Input: Training samples P=$[p_1, p_2, p_3, \ldots, p_k]^T$
Class Label Q=$[q_1, q_2, q_3, \ldots, q_k]^T$
Output: Feature ranked list r

1. Begin with Subset of surviving features **s_set=** [1,2,…n]
2. Feature's ranked list r=[]
3. Repeat until **s_set**= []
4. Restrict training samples to good samples:
   $P = P_o(:,s\_set)$
5. Train the classifier α = ***SVM-train*** (P,q)
6. Calculate weight vector of each feature in s
   $w= \sum_k \alpha_k q_k p_k$
7. Sort weight vector:**sort(w)**
8. Find the features with the smallest weight:
   f=**feature_remove_threshold(w)**
9. Update feature ranked list by removing features below the thresholdfeature
   r = [**s_set** (f), **r**]
10. Eliminate the feature with smallestweight.
    s_set= s_set(l: f – l)

IV. EXPERIMENTS ANDRESULTS

The experiment analysis is done on well know dataset available in public repository is shown in Table 1. All the dataset used for experiment analysis is a two-class linear data consisting of only numerical values as gene expression. The information regarding the dataset used for conducting the experiment is given below:

TABLE 1: DATASET FOR ANALYSIS

Dataset	Total Samples	Number of genes	Class Label
Leukemia	72	3571	ALL AML
Colon	62	2000	Cancer Normal

The experiments conducted mainly based on Leukemia dataset with 3571 genes and 72 samples [14] and colon dataset with 2000 genes and 62 samples [15]. The experiment on the dataset is conducted and the result is analyzed and interpreted based on many factors. Then the cross-validation is done based on an evaluation to analyze the result. The classifier is evaluated using the measures accuracy, error rate, sensitivity and specificity from the confusion matrix as shown in Table 2.

TABLE 2: CONFUSION MATRIX

True True Negative(TN)	False False Positive(FP)
False False Negative(FN)	True True Positive(TP)

Accuracy measure specifies how many samples with a sample a type is predicted as such. The error rate is what percentage of samples are miss-classified. In our dataset numbers of samples are very low and also the number of samples with cancerous tissue is very less. So in order to evaluate such classifier sensitivity and specificity measure is used. Sensitivity is what percent positive samples are classified as positive and the specificity measures what percentages of negative samples are classified as negative.

Firstly the result is analyzed on the existing methods in feature selection and the accuracy is compared with our methodology. And secondly, the threshold selection method is done based on which the number of features is selected. This threshold selection method is analyzed and found the number of features selected and how it impacts the classification process.

The Table 3 shows the accuracy of the classifier without feature selection method.

TABLE 3: ACCURACY WITHOUT FEATURE SELECTION

Dataset	Total Features	SVM Accuracy
Leukemia	3571	87±3
Colon	2000	88±2

Selecting out the relevant features from the dataset is helpful to increase the classification accuracy of the algorithm. The accuracy of the SVM classifier after performing ISVM-RFE is shown in Table 4.

TABLE 4: ACCURACY WITH ISVM-RFE METHOD

Dataset	Total Features	Selected Features	ISVM-RFE Accuracy
Leukemia	3571	3072	97±3
		2572	97±2
		2072	96±2
		1572	97±3
		1072	94±2
Colon	2000	1500	96±2
		1000	96±2
		500	96±2

From the experiments, it is able to prove that the classifier is able to perform well under the feature classification method. Also compared with the existing methodology our result outperforms the other. And in turn, ISVM-RFE decreases the computation time.

REFERENCES

[1] Kerr, M. Kathleen, Mitchell Martin, and Gary A. Churchill. "Analysis of variance for gene expression microarray data." Journal of computational biology 7.6 (2000):819-837.

[2] Guyon, Isabelle, et al. "Gene selection for cancer classification using support vector machines." Machine learning 46.1-3 (2002): 389422.

[3] Kavitha, K. R., Ajith Gopinath, and Midhun Gopi. "Applying improved svm classifier for leukemia cancer classification using FCBF." Advances in Computing, Communications and Informatics (ICACCI), 2017 International Conference on. IEEE,2017.

[4] Thushara, M. G., M. S. Krishnapriya, and Sangeetha S. Nair. "A model for auto-tagging of research papers based on keyphrase extraction methods." Advances in Computing, Communications and Informatics (ICACCI), 2017 International Conference on. IEEE,2017.

[5] Ding, Chris, and Hanchuan Peng. "Minimum redundancy feature selection from microarray gene expression data." Journal of bioinformatics and computational biology 3.02 (2005):185-205.

[6] Awada, Wael, et al. "A review of the stability of feature selection techniques for bioinformatics data." Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on. IEEE,2012.

[7] Ang, Jun Chin, et al. "Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection." IEEE/ACM transactions on computational biology and bioinformatics 13.5 (2016):971-989.

[8] Yin, Jiapeng, et al. "Improving the performance of SVM-RFE on classification of pancreatic cancer data." 2016 IEEE International Conference on Industrial Technology (ICIT). IEEE,2016.

[9] Zhang, Ying, et al. "An efficient feature selection strategy based on multiple support vector machine technology with gene expression data." BioMed research international 2018(2018).

[10] Ani, R., et al. "Modified Rotation Forest Ensemble Classifier for Medical Diagnosis in Decision Support Systems." Progress in Advanced Computing and Intelligent Engineering. Springer, Singapore, 2018. 137-146.

[11] Li,Zifa, WeiboXie, and Tao Liu. "Efficient feature selection and classification for microarray data." PloS one 13.8 (2018):e0202167.

[12] Kavitha, K. R., G. SyamiliRajendran, and J. Varsha. "A correlation based SVM-recursive multiple feature elimination classifier for breast cancer disease using microarray." 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE,2016.

[13] Sahu, Barnali, SatchidanandaDehuri, and AlokJagadev. "A Study on the Relevance of Feature Selection Methods in Microarray Data." The Open Bioinformatics Journal 11.1(2018).

[14] Golub, Todd R., et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." science 286.5439 (1999):531-537.

[15] Dai, Jian J., Linh Lieu, and David Rocke. "Dimension reduction for classification with gene expression microarray data." Statistical applications in genetics and molecular biology 5.1(2006).