International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 3 (2015) pp. 5537-5550
© Research India Publications
http://www.ripublication.com

An Accelerated Feature Subset Selection Using Discriminative Function Allied With Large Margin Distribution Based Ranking

[1]G.Victo Sudha George, [2]Dr. V.Cyril Raj

[1] Research Scholar and Asst.Professor,
Department Of Computer Science and Engineering,
[2] Dean Engineering and Technology,
Dr.M.G.R Educational and Research Institute University,
Chennai-600095, Tamil Nadu, India
Email : [1]sudhajose72@gmail.com

Abstract

The feature selection for classification is a dynamic research area in data mining and optimization. Among the existing methods the effectiveness of the SVM-RFE algorithm is generally considered remarkable, but the prime difficulty in using it is the amount of computational power. The problem becomes worst when bootstraps are needed to improve the feature selection constancy. In this paper, to reduce the computational complexity of the SVM-RFE a new method is proposed that has two important properties to reduce the time of computation. Firstly, in comparison with the conventional SVM-RFE the proposed method uses only the support vectors for training process not all the training samples which reduce the computational time for building a single SVM classifier drastically. Secondly, a new feature ranking criteria , which takes into account both the discriminant ability of individual features along with the ability to maintain the Margin as large as possible even after the dimensionality reduction is proposed to effectively eliminate the irrelevant features and to improve the generalization ability of the classifier. As a result, the performance of SVM-RFE is improved in terms of computational time and selection of highly informative features. This intern improves the generalization ability of the SVM classifier built. The proposed method is tested on Micro array data sets to exhibit its effectiveness.

Key Words : Feature Selection,SVM-RFE,Discrimant ability,Margin Distribution.

1. Introduction

Reduction of feature dimensionality is of considerable importance in machine learning to reduce the computation complexity and to improve the classifier's generalization ability. The task of feature selection consists of selecting a reduced set of features that is most predictive for a given outcome. Such type of problem is found in almost all areas of machine learning research like classification, regression etc. In the context of classification, feature selection aims at finding a subset of features that result in more accurate and compact classifiers. Typically, the error of the designed classifier will drop as more features are added and after reaching some optimal number for a given sample size, it will begin to rise. Therefore the feature selection method must filter out those inputs that are irrelevant to the specific model. Moreover, the selection made should capture the relevant information as well as avoid over fitting the data. In the recent years feature selection has become the focus of numerous research studies in areas where datasets with thousands of features like Gene selection from Micro array data.

It is noticed that the feature selection algorithms generally selects largely varied subsets of features even for small changes to the training data, although majority of these subsets are equally good in terms of classification performance [1**].** Such instability brings down the confidence of domain experts in experimentally validating the selected features. This may build serious issues in applications where feature selection is used as a knowledge discovery tool like Microarray data analysis [2]. So, in gene selection it is important to correct for the selection bias by performing a cross-validation or applying a bootstrap external to the selection process [3][4**]**which further increase the computational complexity of the feature selection process**.**

Though there are many approaches for reducing the size of the feature set, due to it's successful use in selecting informative features, SVM-RFE gained popularity [5,6,7,8]. SVM- RFE, a backward elimination method, ranks all the features according to some ranking function and eliminates a feature with least score. This elimination task is continued until the highest classification accuracy is obtained. The fundamental factors affecting the computational complexity of SVM-RFE are its backward elimination nature and its greediness in selection.Theoriticaly searching backwards from the full set of features available may capture interesting features more easily than the forward selection method, the method is comparatively computationally expensive as it retrains SVM after removing each feature. This is obviously a heavy burden when the sample features are in the thousands. Moreover as finding an optimal subset of features is combinatorially prohibitive, SVM-RFE reduces the complexity of feature selection by being greedy .That is once a feature is selected for removal it is never reintroduced, which intern may affect stability of SVM-RFE method. So, bootstrapping or K-fold cross validations are needed to improve the feature selection stability [9] or to predict the generalization ability without introducing bias [10] which further increases the computational complexity. It is thus desirable to discover ways to speed up the ranking procedure in these cases.

In this paper, we present a more efficient version of SVM-RFE to reduce the computational time of feature selection and improve the generalization ability of the

SVM classifier built . The two properties introduces for improving the efficiency of the conventional SVM-RFE are

1. Concentrate more on the samples (Support Vectors) that are close to the hyperplane which are likely to be misclassified after dimensionality reduction.
2. Try to eliminate more number of irrelevant features in each iteration before retraining the SVM to reduce the computational time. Use the class discriminative ability along with the ability of maintaining the margin as large as possible after dimensionality reduction of the features as the new criteria for ranking the features to improve the generalization ability of the classifier built.

The rest of the paper is organized as follows. Section 2 explains the related works find in the literature followed by Section 3 gives a brief introduction of SVM and Section 4 explains the proposed method in detail. Experimental results are given in Section 5 followed by conclusion and discussion in Section 6.

2. Related Work

Though SVM-RFE is a dominant method for feature selection [11, 12] obtaining an optimal subset of features is combinatorially prohibitive. Most studies have found RFE to choose very informative feature subset but with 12000 or more features to pick from. When the number of samples is more RFE takes substantial computation time. Moreover, RFE is extremely computationally expensive when only one least useful feature is removed during each iteration. To reduce the computational complexity of SVM-RFE aiming at removing more number of features at each iteration before retraining the SVM many approaches were proposed. On this end, a method called E-RFE was proposed to eliminate different numbers of features according to the different weight distributions of a trained SVM [13].The observation is that different combinations of informative features and irrelevant features have different weight distributions. In addition to guarantee the accuracy of the top most ranked features, E-RFE eliminate one feature in each iteration when the numbers of remaining features go below a threshold value. Calculation of three thresholds and slow down the process in case of large number of features are the demerits of this method.

Another method RFE-annealing uses the annealing schedule idea to remove a large number of features in the initial iterations when it is effortless to identify insignificant features and less number of features are removed in the later iterations so that the important features are not removed. In general RFE-Annealing allows an enormous increase in the efficiency of the algorithm without a decrease of classification accuracy. It is made known in most of the cases that when the number of features ranges from one to 200 ,RFE-Annealing gives the same performance as the conventional RFE and also is more stable . But when the number of features selected is around 10 RFE give higher prediction rate than RFE-Annealing [14].

The aim of the proposed method is to provide a more flexible elimination mechanism by considering two aspects. First is to focus on the samples within a narrow region around the separation hyperplane that have the chance of misclassified after dimensionality reduction, Which reduce the time of building the SVM classifier drastically. Secondly, trying to remove more number of features in each iteration . Since the hyperplane that separate the samples of the two different classes is a good indicator of the generalization ability of a trained classifier, we focused on eliminating the features whose removal have less or no effect in the drift of the hyperplane also have less class discriminative power before retraining the SVM classifier.

3. Support Vector Machine

Vladimir Vapnik invented Support Vector Machines (SVMs) in 1979 [15]. They are efficient enough to handle very large-scale classification in both number of samples and number of variables [16]. A special property of SVMs is that they simultaneously minimize the empirical classification error and maximize the geometric margin hence they are known as maximum margin classifier [17] .SVMs are generated in two steps. First, the data vectors are mapped to a high-dimensional space. Second, the SVM tries to find a hyperplane in this new space with maximum margin separating the classes of data. The margin denotes the distance from the boundary to the closest data points (Support Vectors) in the feature space as shown in figure 1.

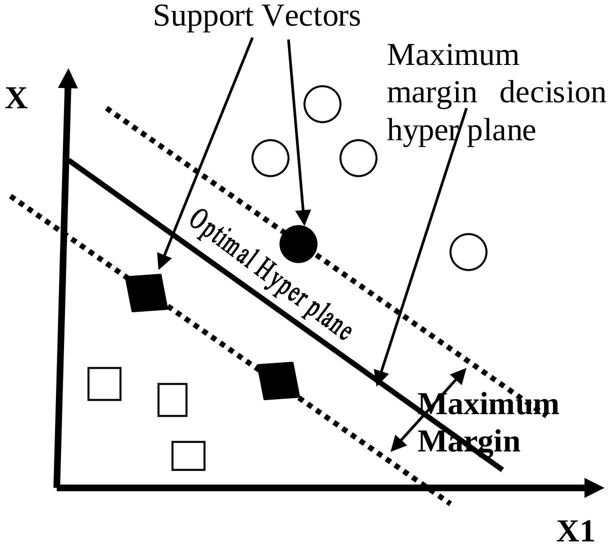


Figure 1 Support Vector Machine

A highlighting feature of SVM is that it is based on the structural risk minimization induction principle, which provides a guaranteed bounded risk value even when the number of the training set is small. But SVM may perform poorly when there are many irrelevant features and for such a situation, feature selection is the remedy.

We are paying attention to linear SVMs because of the kind of the data sets under analysis. A brief review of SVM is given below.

Consider *N* pairs of training samples:

{X(1), Y (1)}, {X(2), Y (2)}, . . . , {X(N), Y(N)},
Where
$X(i) = [x_1(i)x_2(i) \ldots x_k(i)]^T$

is a k-dimensional feature vector representing the i^{th} training sample, and $Y(i) \in \{-1, 1\}$ is the class label of X(i).

A hyperplane in the feature space can be described as the equation

$$W \cdot X+b=0, \tag{1}$$

where $W =[w_1w_2 \ldots w_k]^T$ and b is a scalar. To find the optimal hyperplane one has to solve the following quadratic programming problem: minimize the functional

$$\phi(w) = \frac{1}{2}(w.w) \tag{2}$$

under the inequality constraints

$$y_i[(X_i \cdot W)+b] \geq 1, i = 1,2,\ldots\ldots\ldots, N \tag{3}$$

We can obtain the functional

$$W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum \alpha_i \alpha_j y_i y_j x_i^T x_j \tag{4}$$

It remains to maximize this functional under the constraint

$$\sum \alpha_i y_i = 0, \alpha_i \geq 0, i = 1,\ldots.,N \tag{5}$$

Once the optimization problem has been solved, we can obtain w w as follows:

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \tag{6}$$

It is usually the case that most of the parameters α_i are zero. The decision hyperplane therefore only depends on a smaller number of data points with non-zero α_i; these data points are called support vectors. So we can change the equation (6) as

$$w = \sum_{i \in SV} \alpha_i y_i x_i \tag{7}$$

As a result, equation (2) can be obtained and the SVM classifier has been built.

4. Proposed Method

Hypothesis used in the Proposed Method

By investigating the characteristics of SVMs, it can be found that the optimal hyperplane and support vectors of SVMs can be used as measures of the significant subset of features. Therefore, these measures can be used to select the most informative features.

Hypothesis 1: Concentrate more on the samples (Support Vectors) that are close to the hyperplane which are likely to be misclassified after dimensionality reduction.

Since the training samples other than support vectors have no contribution to determine the decision boundary and as the number of support vectors is relatively small, we could train SVM just by using the support vectors. Following this idea, we propose the modified version of SVM-RFE algorithm, which learn SVM discriminative function using all features of all the training samples at first ie. is only once. In the following iterations an active subset as the candidates of the support vectors is dynamically maintained, and trains SVM using this reduced subset rather than the entire training set. In this way, we are able to find the boundary with less computational time.

Hypothesis 2: Try to eliminate more number of irrelevant features in each iteration before retraining the SVM to reduce the computational time. Use the class discriminative ability along with the ability of maintaining the margin as large as possible even after dimensionality reduction of the features as the new criteria for ranking the features to improve the generalization ability of the classifier built.

The original SVM-RFE uses the squared coefficients w^2 of the weight vector w obtained from Eqn (2) as feature ranking criteria. By instinct, those features with the largest weights are the most informative[18]. Thus in an iterative procedure of SVM-RFE ,the weight vector of all the features is computed by training the SVM classifier, and eliminate the feature having the least weight . This process is continued until a small subset of features is obtained. However, this way for feature ranking is a greedy method and we should look for more evidences for feature selection. From the observations in [19, 20, 21,22], we conclude that the largest margin is a good measure of the generalization ability of a trained classifier. It is noticed that by construction, an SVM classifier assert on a fat margin around the decision boundary. And Therefore features whose removal does not cause drifting or shrinking of the margin of a trained SVM represent significant discriminative power (performance difference of the SVM before and after removing these features is expected to be little).However, because the dimensionality of original input space has been reduced after feature selection, the margin is always to reduce in size and what we can do is trying our best to make the shrink small to some extent. Therefore, in feature selection process, we should prefer the features which make more contribution to maintaining the margin large. This is an additional evidence for feature ranking. To comprehend this idea, we bring in a coefficient CLM_k (Coefficient of Large margin)

$$CLM_k = \frac{1}{S_{n_+}} \sum_{i \in S_+} x_{i,k} - \frac{1}{S_{n_-}} \sum_{j \in S_-} x_{j,k} \tag{8}$$

$$n = S_{n+} + S_{n-} \tag{9}$$

Where S_+ denoted the Support Vectors belong to positive samples, S_- denotes the Support Vectors belong to negative samples and n denotes total number of Support Vectors, S_{n+} denotes the number of S_+ and $S_{n\_}$ denotes the number of S_-, and $x_{i,k}$ denotes the kth feature of Support Vector i in the input space. The larger CLM_k indicates that the kth feature of input space can make more contribution to maintaining the margin large. Therefore, CLM_k can assist $|w_k|$ for feature ranking. The solution is that, combining the two evidences, we can order the features by ranking $CLM_k|w_k|$.

The Proposed Modified SVM- RFE Feature Selection Algorithm

We consider N training samples with k features each and their corresponding class labels. Initially train the SVM using all the features of all the training samples like the conventional SVM-RFE. After training SVM with these N sample and label pairs, we find the discriminative function value $f(X_i)$ for the Support Vectors and optimal weight vector W of each feature. And for improving the generalization ability of the classifier we compute the parameter CLM_K for all the features of support vectors. Since $y_i f(X_i)$ of the Support vectors control the Maximum Margin Distribution of the SVM calculate it for all the Support Vectors and store it in a linear Vector MMDV. The proposed method now keep on removing features as long as it is not drifting the margin which is determined by checking with the pre-defined threshold limit. When it goes beyond the threshold SVM is retrained. To avoid losing important feature, the algorithm will automatically start eliminating one feature at a time when the number of features reaches a specified number. The outline of the algorithm is given below.

1. Initialization.
   An empty ranked feature list $RANK_{list.}$
   A set of features to be ranked $F = \{1,2,\ldots\ldots,K\}$
2. Set $l = K/2$
3. Train the SVM with K features of the N training sample set and obtain its optimal weight vector
   $W^{[K]} = [w_1, w_2, w_3, \ldots\ldots, w_K]$,$CLM^{[K]} = [CLM_1, CLM_2, \ldots\ldots, CLM_K]$ and bias $b_{k.}$
4. Rank the K features in the descending order of $W(i) \bullet CLM(i)$, $i = 1,2,\ldots.,K$.
5. Compute discriminative function for the Support Vectors using
   $f(X_i^{[k]}) = < W^{[k]}, X_i > + b_K, i = 1,2,\ldots\ldots,N$ Where $\alpha_i > 0$
   The Margin Distribution of the SVM MD_K^n using
   $y_i \ f(X_i^{[k]}),\ i = 1,2,\ldots\ldots,N$ Where $\alpha_i > 0$
6. Set $W^{[K-l]} = [w_{l+1}, w_{l+2}, w_{l+3}, \ldots\ldots, w_K]$,$CLM^{[K-l]} = [CLM_{l+1}, CLM_{l+2}, \ldots\ldots, CLM_K]$
   and $X_i^{[K-l]} = [x_i^{l+1}, x_i^{l+2}, \ldots\ldots, x_i^K]$, $I = 1,2,\ldots\ldots, n$ where n is the number of support vectors
7. Compute MD_{K-l}^n using $y_i \ f(X_i^{[k-l]})$, $i = 1,2,\ldots\ldots,N$ Where $\alpha_i > 0$
8. Calculate $\Delta MD_k = \| MD_K^n - MD_{K-l}^n \|$
9. If $\Delta MD_k > TH_{\Delta MMD}$, Let $l = l/2$ go to 5

10. Else remove *l* least significant features and place them on the top of $\text{RANK}_{\text{list}}$
11. set K= K-*l*
12. If K> 1 go to step3 else display $\text{RANK}_{\text{list}}$ and Stop.

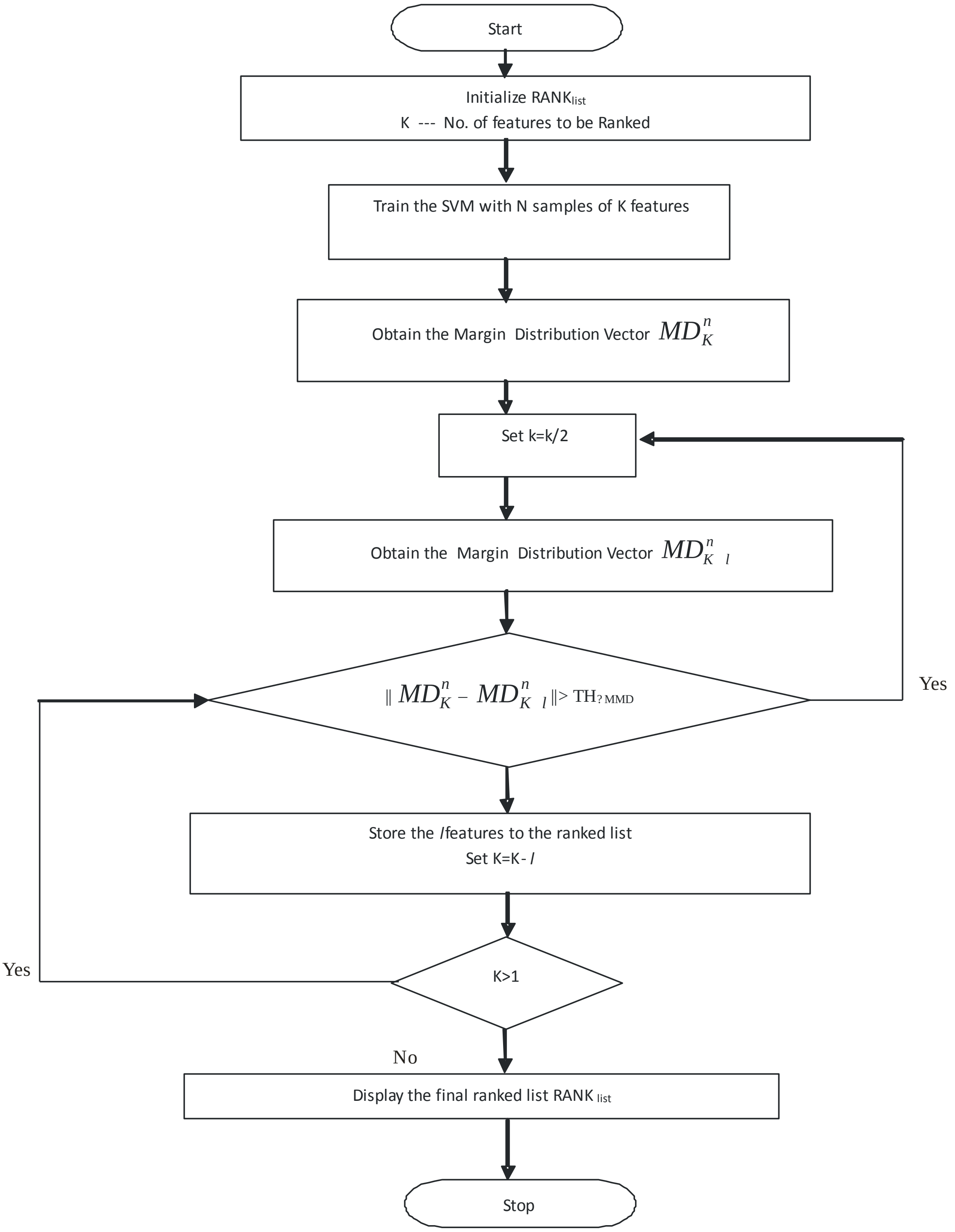


Figure 2 Flow chart of the proposed method

5. Experimental Results

The proposed method is tested for gene selection and classification of DNA microarray data. DNA microarray is a technology that measures expression levels of thousands of genes simultaneously. As microarray data contain thousands of features of which many are irrelevant or redundant we found it as a good candidate to test our proposed method. The details of the three data sets used to test the performance of the proposed method are given in table 1. To reduce the complexity of computation ,an initial filtering was carried out using BRB Array Tool on all the three datasets to filter out the irrelevant noisy genes by setting p<0.001 and 3 fold change . In the following simulations, SVM-RFE, M-RFE and ERFE are coded based on the linear SVM routine of MATLAB . The reported execution time is based on an Intel Celeron CPU of 2.40 GHz.

Table 1. Details of the data sets used

Data Set	No. of Samples	Class1	Class2	No. of genes	Training Set		Testing Set		No. of genes after initial filtering
					Cass1	Class2	Cass1	Class 2	
Breast cancer dataset	86	43	43	22285	32	32	11	11	613
AML-ALL dataset	72	47	25	7129	37	17	10	08	413
Colon cancer dataset	62	22	40	6500	16	30	06	10	389

Note : AML- ALL data set Class1-ALL Class2-AML and other two data sets class1-Normal class2-cancerous

Description of the data sets used in the experiments is given in Table 1. All the three data sets are randomly divided into training set and testing sets. The training sets were used to rank the features and to train the SVMs with different number of ranked features. The testing sets were used to predict the generalization ability of the trained SVMs. Leave One Out Cross Validation (LOOCV) is used to train the SVMs.For example in the case of the breast cancer data set the proposed algorithm was first tested on the 64 training samples. At each of the 64 iterations 63 samples were used for selecting the genes and to train the SVM classifier, and the remaining 1 sample was used to test the gene subset and the classifier built. The classification error is the average of the classification errors on the test samples. In the testing phase genes that were selected from the 64 training samples and the performance was measured on the 22 testing samples.

The Figure 3 gives the average classification error on the testing samples versus various training set sizes for the breast cancer data set. The performance of the proposed method along with the conventional SVM-RFE and other feature selection algorithms E-RFE and RFE-Anealing are also presented for comparison. It is very clear that the proposed method outperformed the other methods especially when the training set is small. As more and more number of training samples are used SVM-RFE also performed well like the proposed method. But comparing the other two methods E-RFE and RFE-Anealing which also removed more number of features in each iteration of feature selection the proposed method gives appreciable generalization ability. It is very good indication that the criteria which the proposed method used for feature selection, combination of class discrimination ability and the ability of the feature to maintain the margin as large as possible even after dimensionality reduction has improved the generalization ability of the classifier. On the other two data sets also similar results are observed.

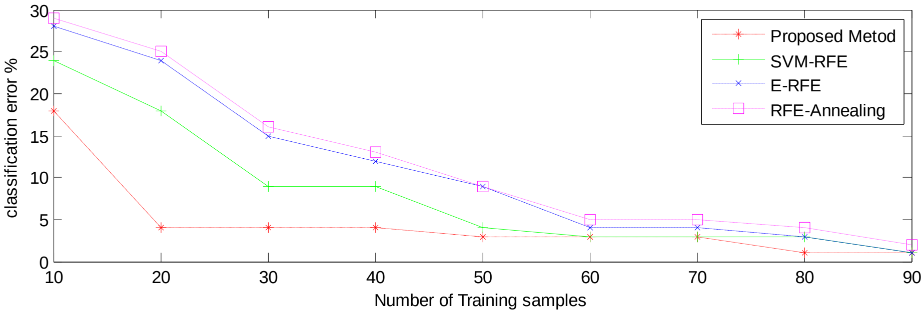
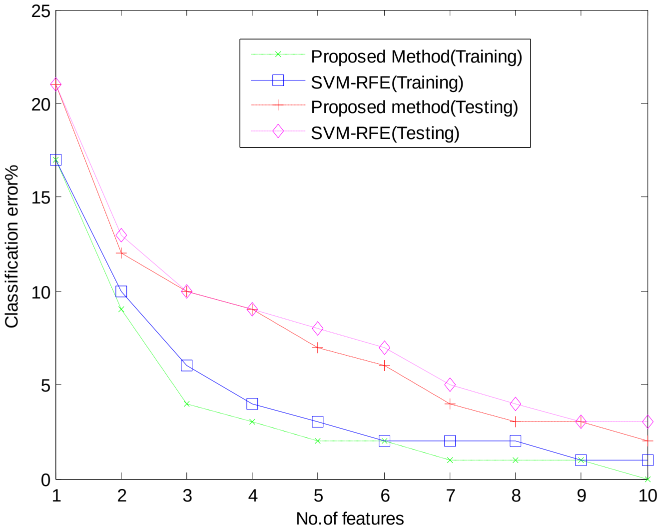
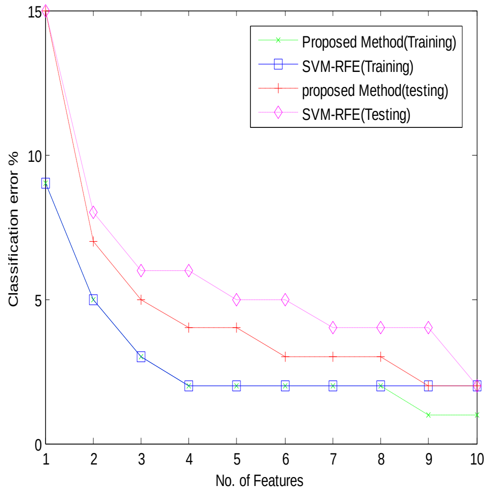


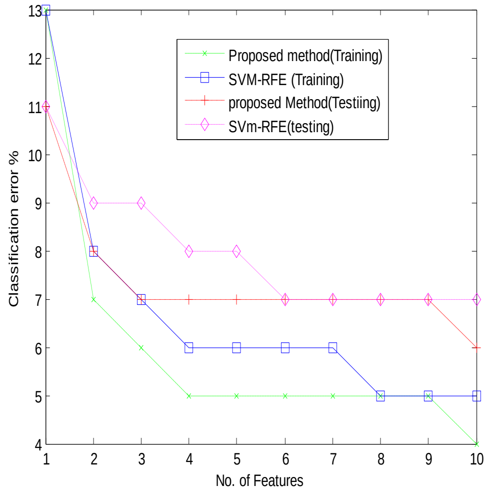
Figure 3 Comparison of feature selection methods on breast cancer data set



(a) Breast Cancer data set



(b) AML-ALL Data set



(C) colon Cancer data set

Figure 4 The comparison of the generalization ability of the proposed method with SVM-RFE

Figure 4 shows the performance comparison on the three data sets. Throughout the experiments the proposed method has shown better generalization ability compared to and the original SVM-RFE .This is a good indication for the ranking criteria proposed by us. It is also seen from the table that in case of all the three data sets speed up gained by the proposed method is more than 50 %. More detailed results are listed in Table2, which evidentially shows that the proposed method improves the efficiency of SVM-RFE without sacrificing the accuracy of either feature selection or the classification.

Data Set	Classification accuracy(%)				Run Time(s)		
	Training		Testing		Proposed Method	SVM-RFE	Speed Up %
	Proposed Method	SVM-RFE	Proposed Method	SVM-RFE			
Breast cancer dataset	98.8	97.4	94.9	93.6	15.5	24.2	64.0
AML-ALL dataset	97.4	96.4	96.3	95.4	10.6	13.3	79.7
Colon cancer dataset	95.3	94.1	92.3	91.2	19.7	33.9	58.1

Table 2 Performance Comparison between the proposed method and SVM-RFE

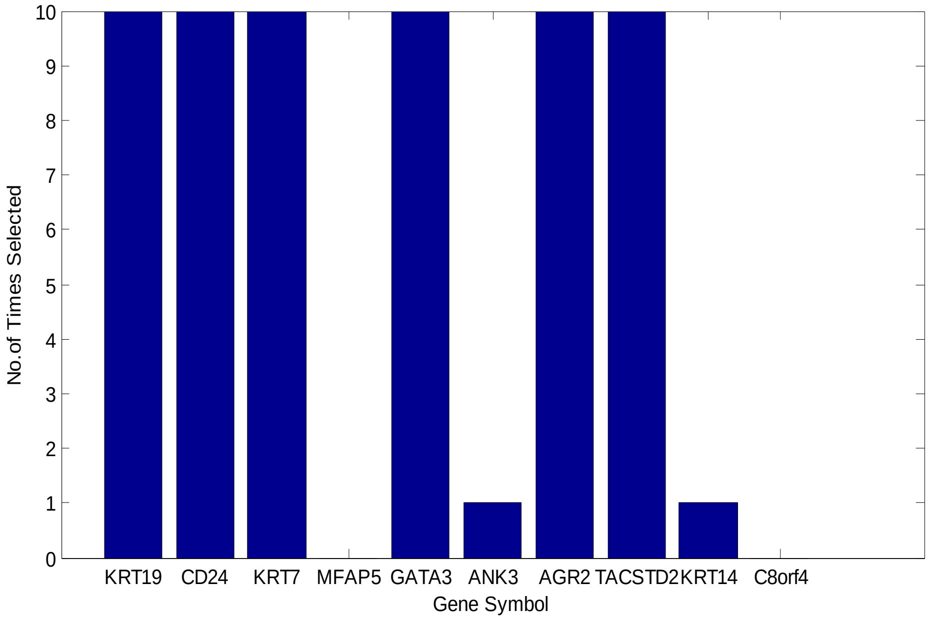


Figure 5 Stability of the proposed method on Breast cancer dataset

To verify the stability of the proposed method experiment is conducted for the breast cancer dataset. Figure 5 shows the total number of times each feature has been selected over 10 trials for the breast cancer dataset. The figure is plotted for Gene symbol versus number of times a gene being selected. From the figure it is seen that the genes with symbols KRT19 ,CD24, KRT7, GATA3, AGR2 TACSTD2 are selected all the 10 times where as ANK3 and KRT14 are selected one time and C8orf4 MFAP5 are not at all selected in the 10 trials. This clearly indicates that the proposed method gives very good stability in terms of feature selection also.

6. Conclusion

In this paper, we present an accelerated feature selection method which is an enhancement of SVM-RFE. First an initial filtering was done to remove the noninformative features. As a result the number of SVM training is reduced. To make the feature pruning process more effective, we formulated a new feature ranking criterion to take into account the class discriminative power of the features along with the ability of the features to retain the size of the margin as large as possible even after feature elimination. The results of the experiments conducted on the three microarray data sets demonstrated that our feature ranking criteria has improved the generalization ability of the classifier built from the selected features. Furthermore, during the feature selection process, an active training set is maintained as the candidates of the support vectors. This reduced the training set of SVM considerably. In this way, the number of samples participating in a single time SVM construction decreases, and the training process is ultimately speeded up. It is demonstrated that the proposed method has reduced the computational time as well as improved the generalization ability of the SVM-RFE algorithm.

7. References

[1] Kalousis, J. Prados, and M. Hilario, 2007 ,"Stability of feature selection algorithms: a study on high-dimensional spaces," Knowledge and Information Systems, vol. 12, pp. 95–116.

[2] M. S. Pepe, R. Etzioni, Z. Feng, et al., 2001 ,"Phases of biomarker development for early detection of cancer," J Natl Cancer Inst, vol. 93, pp. 1054–1060.

[3] C. Ambroise and G. J. McLachlan, 2002, "Selection bias in gene extraction on the basis of microarray gene-expression data," Proc. Nat. Acad. Sci. USA, vol. 99, pp. 6562–6566.

[4] A. D. Baxevanis and B. F. F. Ouellette, 2001,Bioinformatics. New York:Wiley.

[5] Fernandez-Lozano C, Fernández-Blanco E, Dave K, Pedreira N, Gestal M, Dorado J, Munteanu CR. ,2014,"Improving enzyme regulatory protein classification by means of SVM-RFE feature selection",10(5):1063-71

[6] Pember A. Mundra and J. C. Rajapakse, 2007 ,"SVM-RFE with relevancy and redundancy criteria for gene selection," in PRIB, J. C. Rajapakse, B. Schmidt, and L. G. Volkert, Eds., vol. 4774, Springer, , pp. 242–252.

[7] Tang, Y.-Q. Zhang and Z. Huang, 2007 ,"Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis", IEEE/ACM Trans. Comput. Biology Bioinform, vol. 4, no. 3, pp. 365–381.

[8] K. Duan and J. C. Rajapakse, 2005, "Svm-rfe peak selection for cancer classification with mass spectrometry data," in APBC, pp. 191–200.

[9] I. Guyon and A. Elisseeff, 2003 , "An introduction to variable and feature selection," J. Mach. Learn. Res., vol. 3, pp. 1157–1182.

[10] M. Dash and H. Liu, 1997, “Feature selection for classification,” Intell. Data Anal, vol. 1, no. 1-4, pp. 131–156.

[11] Guyon I, Weston J, Barnhill SMD, Vapnik V2002,” Gene Selection for Cancer Classification using Support Vector Machines.Machine Learning “, 46(1–3):389-422

[12] Furlanello C, Serafini M, Merler S, Jurman G 2003, “Entropy-based gene ranking without selection bias for the predictive classification of microarray data”, BMC Bioinformatics.

[13] C.Furlanello,M.Serafini,s.Merler, and G. Jurman, 2003,”An accelerated procedure for the recursive feature ranking on microarray data, “Neural Networks,Vol.16,No.5-6,pp.641-648,June –July.

[14] Yuanyuan Ding and Dawn Wilkins, 2006,” Improving the performance of SVM-RFE to select genes in Microarray Data,Bioinformatics ,7(Suppl 2):S12.

[15] Vapnik V, 1995,The Nature of Statistical Learning Theory Springer-Verlag;.

[16] Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S, 2005, “A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis”, Bioinformatics 21(5):631-643.

[17] Wikipedia: Support vector machine.

[18] Guyon, I., Weston, J., Barnhill, S., Vapnik, V. N., 2002, Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning, vol. 46, 389-422.

[19] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, 1998 ,“Boosting the margin: a new explanation for the effectiveness of voting methods,” The Annals of Statistics, vol. 26, no. 5, pp. 1651–1686.

[20] J. Shawe-Taylor and N. Cristianini, 1999, “Further results on the margin distribution,” in Proceeding of 12th Annual Conference on Computer Learning Theory, Newyork, NY, ACM Press.

[21] L. Mason, P. L. Bartlett, and J. Baxter, 2000 ,“Improved generalization through explicit optimization of margins,” Machine Learning, vol. 38, no. 3, pp. 243–255.

[22] A. Garg and D. Roth, 2003 ,“Margin distribution and learning algorithm,” in Proceeding of 20th International Conference on Machine Learning, Washington DC.