Hybridized KNN and SVM for gene expression data classification☆

Zhen Mei, Qi Shen*, Baoxian Ye

Chemistry Department, *Zhengzhou University*, *Zhengzhou*, *Henan* 450052, *China*

Received October 22, 2008

Abstract

Support vector machine (SVM) is one of the most powerful supervised learning algorithms in gene expression analysis. The samples intermixed in another class or in the overlapped boundary region may cause the decision boundary too complex and may be harmful to improve the precise of SVM. In the present paper, hybridized k-nearest neighbor (KNN) classifiers and SVM (HKNNSVM) is proposed to deal with the problem of samples in the overlapped boundary region and to improve the performance of SVM. The first KNN is used to prune training samples and the second KNN is combined with SVM to classify the cancer samples. The proposed algorithm was used in binary and multiclass classification of gene expression data. The results were compared to those obtained by single SVM and KNN. It has been demonstrated that the proposed method is a useful tool for classification and the misclassification rate for the prediction set is reduced with samples pruning used. Compared with SVM and KNN, the misclassification rates of HKNNSVM for the datasets containing mislabeled samples were notably lower than that by SVM and KNN, which indicated that the classification performance of HKNNSVM was stable. [Life Science Journal. 2009; 6(1): 61 – 66] (ISSN: 1097 – 8135).

Keywords: support vector machine; k-nearest neighbor; gene expression data; classification

1 Introduction

Nowadays, people can obtain the expression datasets of thousands of genes simultaneously using microarray technology. One of the important fields in using these gene expression datasets is to classify and predict the diagnostic category of a sample[1,2]. Actually, precise diagnosis and classification is crucial for successful treatment of illness.

For classifying microarray data, one can use the classical liner discriminant analysis, artificial neural networks, KNN, as well as some more sophisticated machine learning methodologies including bagging, boosting and kernel methods. Among them, SVM is one of the most powerful supervised learning algorithms in gene expression analysis. SVM has been found generalization ability and useful in handling classification tasks in case of the high dimensionality and sparsity of data points.

SVM constructs an optimal hyperplane from a small set of samples near the boundary and is sensitive to these boundary samples. The samples intermixed in another class or in the overlapped boundary region may cause the decision boundary too complex and may be harmful to improve the precise of classifier. The existence of samples in the overlapped region may also increase the computation burden and decline the generalization ability of classifier. In addition, labeling a sample in some cases can be subjective and a few mislabeled samples could deeply degrade the performance of the classifier[3]. Mislabeled and troublesome learning samples may be often near the boundary and lead to a result with high error rate. Many researches[4,5] have been focused on identifying and pruning the questionable redundancy samples to improve the performance of classification.

There is increasing evidence that the ensemble classifier performs better than the individual. The combined classifiers increase not only the accuracy of the classification, but also lead to greater confidence in the result[6]. Though SVM has been found useful in handling classification tasks, it has been recognized that results of

☆Supported by the National Natural Science Foundation of China (Grant No. 20505015, 20475050).

*Corresponding author. Tel: 86-371-67781024; Fax: 8637167763220; Email: shenqi@zzu.edu.cn

SVM analysis can be improved when combining with other classifiers.

In the present paper, hybridized KNN and SVM (HKNNSVM) is proposed to deal with the problem of samples in the overlapped boundary region and to improve the performance of SVM. KNN[7] is a very efficient pattern recognition method and can be easily carried out. In a statistical opinion, the error rate of a KNN classifier tends to the Bayes optimal when k and the size of sample set tend to infinity[8]. Base on these advantages, KNN is introduced into SVM to classify three gene expression datasets. We firstly used the KNN to prune training samples and then combine KNN with SVM to improve the classification. The proposed hybridized algorithm was used in binary and multiclass classification of gene expression data. The results were compared to those obtained by single SVM and KNN. In this study, linear kernel function is included in the SVM and HKNNSVM procedure, so the SVM, KNN and HKNNSVM are linear process. It has been demonstrated that the proposed method is a useful tool for classification and the classification performance is stable. It has indicated that the proposed classifier is superior to some other classifier.

The remainder of this paper is organized as follows: In Section 2, we provide the detail of our proposed procedure. Section 3 introduces three public datasets to evaluate the performance of our proposed method. The experimental resulted from our proposed method is presented, and compared with KNN and SVM method on the public datasets in section 4. Finally conclusions are drawn in section 5.

2 HKNNSVM

SVM is sensitive to these samples intermixed in another class or these boundary samples. The existence of samples in the overlapped region may be harmful to the performance of SVM. The hybridized classifier can improve the precise of classification, so hybridized of KNN classifiers and SVM is proposed to improve the performance of classification. The first KNN is used to prune training samples and the second KNN is combined with SVM to classify the cancer samples. We first yield the distances matrix which is a symmetrical matrix containing the Euclidean distance between each pair of samples. Then the K nearest neighbors for each sample are sought. In the first KNN, if the class label of training sample is same as the label of the majority of its K nearest neighbors, the training sample is reserved, whereas others are pruned. For the pruned samples set, the second KNN and SVM are applied to classify. If k nearest neighbors have all the same labels, the sample is labeled. Otherwise SVM will be applied to classify the rest sample. The hybridized KNN and SVM are described as follows.

Step1. Select relevant genes using t-test. The gene selection is an important aspect for class identification and t-test is one of the most popular gene ranking methods.

Step 2. Prune training samples by the first KNN.

Step 3. Use the second KNN to classify the remaining samples which are not pruned away from the training set. The sample is classified into the same class that its K neighbors are all in the same class, otherwise, go to the next step.

Step 4. Apply SVM to classify the rest unidentified samples. The HKNNSVM scheme is presented in Figure 1.

For test sample, the k nearest objects to it in training dataset are selected firstly, if all the k nearest objects belong to category L, then classify the test sample in L. otherwise, apply the SVM to label it.

In this study, linear kernel function is included in the SVM procedure.

3 Datasets

Three public datasets were used to test our method in this paper.

3.1 Colon data

This dataset[1] is often used for testing all kinds of classification method. It consists of 62 samples (40 tumor and 22 normal colon tissues). Gene expressions for 2000 human genes are measured using the Affymetrix technology. These data are publicly available at http://microarray.princeton.edu/oncology/affydata/index.html. For colon dataset, we constructed 50 randomly selected samples (18 normal and 32 cancer tissues) as training set and the remaining 12 samples as the prediction set.

3.2 Estrogen data

These datasets were obtained by applying the Affymetrix gene chip technology and first presented in papers by West and Spang[14,15]. The common expression matrix monitors 7129 genes in 49 breast tumor samples. In this dataset, 25 samples are labeled ER$^+$, the rest 24 samples are labeled ER$^-$. These data are retrieved from http://mgm.duke.edu/genome/dna micro/work/. Among 49 estrogen samples, 40 randomly selected samples (20

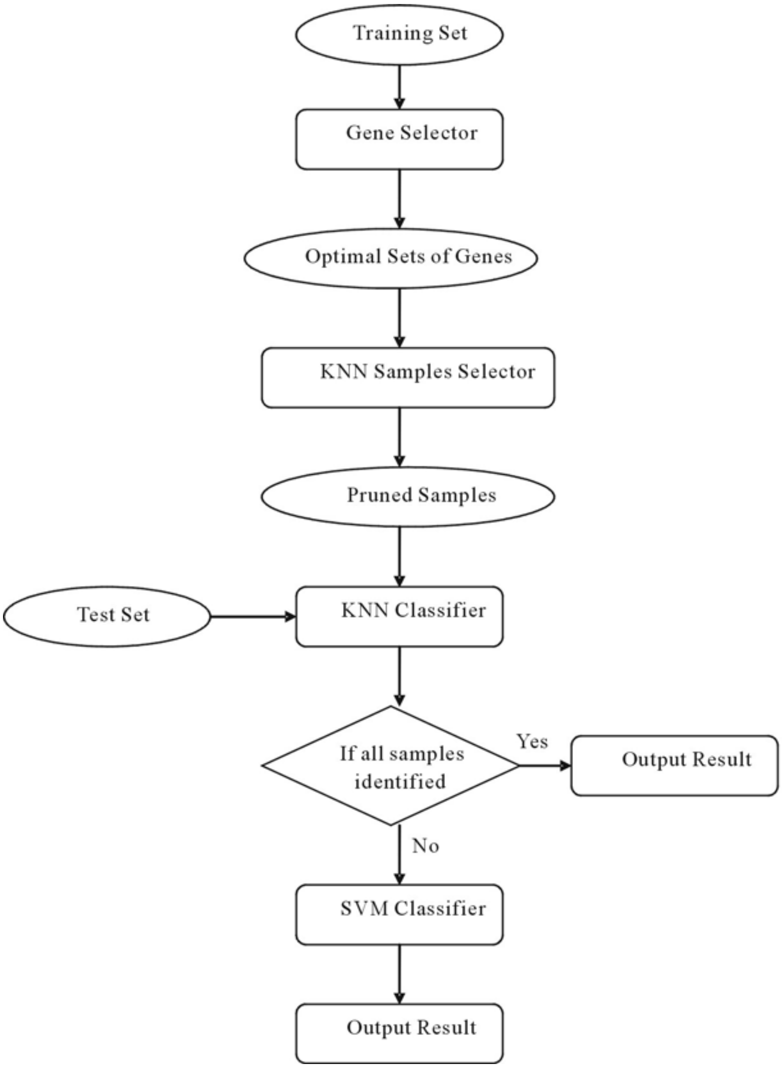

Figure 1. The chart of the HKNNSVM scheme.

are ER$^+$ and 20 are ER$^-$) were used as training set and the remaining 9 samples as the prediction set.

3.3 Acute lymphoblastic leukemia (ALL) data

ALL dataset (16) is publicly available at http://www.stjuderesearch.org/data/ALL1/all_datafiles.html. The dataset consists of expression profiles of 12625 human genes from 248 patients, there are 6 subsets: 15 BCR-ABL samples, 27 E2A-PBX1 samples, 64 Hyperdiploid > 50 chromosomes samples, 20 MLL samples, 43 T-All samples, 79 TEL-AML1 samples. For ALL dataset, we constructed 208 randomly selected samples (five-sixths for each subset) as training set and the remaining 40 samples as the prediction set.

The HKNNSVM algorithm was programmed in Matlab 6.0 and run on a personal computer (Intel Pentium processor 733MHZ 256 MB RAM).

4 Results and Discussion

4.1 The performance of classifiers for three gene expression datasets

In the present study, three publicly available datasets are used to test the performance of our method for tumor classification. The performance of classifier is measured according to an averaged classification error rates over 5-fold cross-validation. Briefly, the samples are randomly split into five data sets of approximately equal size respectively. The training data set, which is four parts of the subsets, is used to derive a classification model that is then applied to predict the remaining subsets. The procedure should be repeated five times and the classifier is evaluated by the averaged error value over the five subsets. Because of the arbitrariness of partition of dataset, the predicted error rate of a model at each iteration is not necessarily the same. To evaluate accurately the relevance of genes subset, such 5-fold cross-validation was repeated 300 times and then averaged the error rate.

As a comparison, support vector machine and K-nearest neighbors classification methods were first utilized for these gene expression datasets. The t-test was first used on the training data to select the optimal genes. For each dataset, we select 30, 50, 100, 200 and 500 top-ranked genes using t-test statistic respectively to test the classification and then selected the best subset that obtained the lowest error. Errors for several fixed feature size for each classifier are showed in Figures 2, 3 and 4.

The misclassification rates of training and test set for each dataset by SVM and KNN were presented in Table 1.

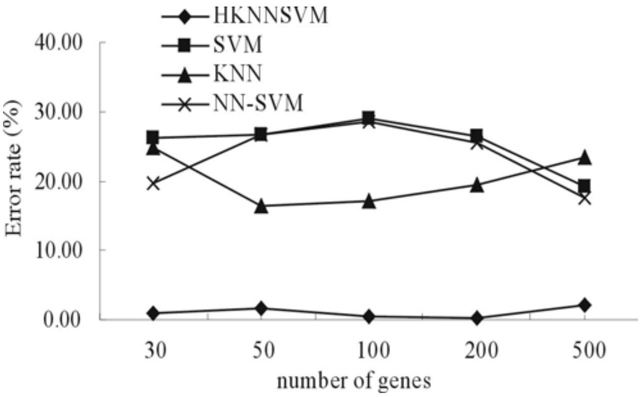

Figure 2. Error rate for colon dataset.

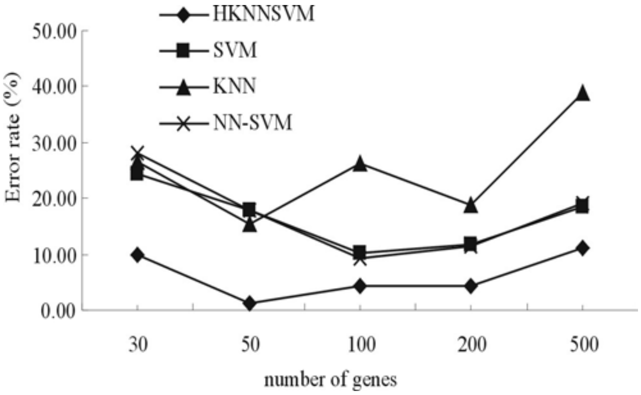

Figure 3. Error rate for Estrogen dataset.

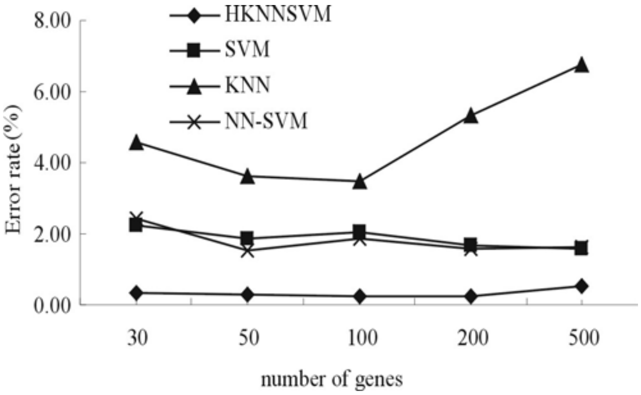

Figure 4. Error rate for ALL dataset.

For colon dataset, the optimal subset for SVM classification contains 500 genes and the misclassification rate for training set and test set were 19.20% and 12.39% respectively. Using KNN the optimal subset contains 50 genes and the misclassification rate for training set and test set were 16.40% and 12.22% respectively. For estrogen and ALL dataset, a comparison with KNN classifier shows that better results were obtained from SVM algorithm. For estrogen dataset, the optimal subset for SVM contains 100 genes. Using all 100 genes, the misclassification rate by SVM for training and prediction sets are 10.13% and 7.94% respectively. Using the optimal 500 genes, the misclassification rate for training and test sets by SVM for ALL dataset were 1.56% and 1.74% respectively.

To compare with HKNNSVM, the KNN combined with SVM classifier (KNNSVM) was also performed in which training samples was not pruned. In KNNSVM, KNN is used only for classification and not to prune training samples. The classification results of KNNSVM were also presented in Table 1. The misclassification rates for the three datasets were 11.97%, 7.85% and 1.75% respectively. A comparison of KNNSVM, SVM and KNN shows that KNN is inadequate for classification of estrogen and ALL data and better results were obtained from KNNSVM and SVM.

To further improve the classification accuracy of the prediction models, the HKNNSVM algorithm was used to evaluate the misclassification rates of the three gene expression datasets. In HKNNSVM, the first KNN is used to prune training samples and the hybridized of the second KNN and SVM is applied to classify the cancer samples. The best classification performances of the KNN are achieved when k takes values from 3 to 5. In the present work, K is selected as 5 in the first and second KNN by experience. The misclassification rates for each dataset by HKNNSVM were also showed in Table 1. From Table 1, one can see that the misclassification rate for colon dataset was 9.75% using 200 genes. For estrogen dataset the misclassification rate was 6.33% using 50 genes. Comparing with SVM, KNN and KNNSVM, the misclassification rates by HKNNSVM are lower than that by SVM, KNN and KNNSVM, and moreover, the number of genes used by HKNNSVM was less. For ALL dataset, using the optimal 200 genes, the misclassification rate for training and test sets by HKNNSVM were 0.23% and 2.20% respectively. The misclassification rate of test set for ALL dataset by HKNNSVM is higher than that by SVM and KNNSVM. The reason for this is probably that there are only 15 samples in one of the subset of ALL dataset (BCR-ABL) and pruned training set may be too few to build classification model.

Table 1. Results of misclassification rates for three datasets

Datasets		Method			
		HKNNSVM	SVM	KNN	NN-SVM
Colon	Number of genes	200	500	50	500
	Training set	0.28%	19.20%	16.40%	17.50%
	Test set	9.75%	12.39%	12.22%	11.97%
Estrogen	Number of genes	50	100	50	100
	Training set	1.26%	10.13%	15.38%	9.25%
	Test set	6.33%	7.94%	12.19%	7.85%
ALL	Number of genes	200	500	100	500
	Training set	0.23%	1.56%	3.46%	1.61%
	Test set	2.20%	1.74%	3.69%	1.75%

4.2 The effectiveness of pruning samples

To check the effectiveness of pruning samples, ten percent of the training sample were selected randomly and mislabeled. We investigate if those mislabeled samples can be eliminated effectively by the first KNN that is used to prune training samples and the performance of classifiers affected by those mislabeled samples.

We selected and mislabeled five and four samples randomly for colon and estrogen dataset respectively. For ALL dataset, not each subset has an overlapped region with others and more than 95% overlapped region arose between BCR-ABL and Hyperdiploid > 50 samples. That is to say, the superposition of BCR-ABL and Hyperdiploid > 50 samples is the most serious one and the key to improve the performance of classifier is to classify samples of the two subsets successfully. So in this study, 6 samples of BCR-ABL and Hyperdiploid > 50 in training set were mislabeled. The percentage of pruned mislabeled samples is computed to evaluate the effective of pruning mislabeled samples. For colon dataset, 86.20% mislabeled samples were pruned. The percentage of pruned mislabeled samples are 88.25% and 97.56% for estrogen and ALL datasets respectively, indicating KNN is effective for pruning mislabeled samples.

Table 2 summarizes the classification results of datasets contained mislabeled samples. Comparing with the results of Table 1, the misclassification rates for the three datasets contained mislabeled samples were increased largely using SVM, KNN and KNNSVM classifier and a very small number of mislabeled samples could deeply degrade the performance of the classifier. Using HKNNSVM classifier, the misclassification rates are 10.78%, 9.74% and 2.03% for colon, estrogen and ALL dataset respectively. Comparing with SVM, KNN and KNNSVM, the misclassification rates by HKNNSVM were notably lower than that by SVM, KNN and KNNSVM. The introduction of pruning training samples into the hybridized of KNN and SVM improved the characteristic performance of the classifier, as the misclassification rate for the prediction set is stable and reduced with samples pruning used.

Table 2. Results of misclassification rates for dataset contained mislabeled samples

Datasets		Method			
		HKNNSVM	SVM	KNN	NN-SVM
Colon	Number of genes	200	500	50	500
	Training set	3.74%	28.7%	19.40%	26.7%
	Test set	10.78%	20.61%	12.86%	18.92%
Estrogen	Number of genes	50	100	50	100
	Training set	5.42%	23.62%	18.50%	17.13%
	Test set	9.74%	19.0%	15.41%	18.26%
ALL	Number of genes	200	500	100	500
	Training set	0.56%	2.48%	4.45%	4.16%
	Test set	2.03%	3.07%	3.53%	4.08%

5 Conclusion

In this paper, we applied the hybridized of KNN and

SVM for gene expression data classification. Our method test three public datasets and have good performance. The proposed method was used to prune the mislabeled training samples and can eliminate those samples effectively. Meanwhile the misclassification rates of our propose method not increase sharply.

References

1. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 1999; 12: 6745 – 50.
2. Shah S, Kusiak A. Cancer gene search with data-mining and genetic algorithms. Comput Biol Med 2007; 2: 251 – 61.
3. Malossini A, Blanzieri E, Raymond,T. Detecting potential labeling errors in microarrays by data perturbation. Bioinformatics 2006; 22: 2114 – 21.
4. Sun J, Hong GS, Wong YS, Rahman M, Wang ZG. Effective training data selection in tool condition monitoring system. Int J Mach Tool Manu 2006; 46: 218 – 24.
5. Angelova A, Abu-Mostafa Y, Perona P. Pruning training sets for learning of object categories. Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2005; 1: 494 – 501.
6. Kim KJ, Cho SB. Ensemble classifiers based on correlation analysis for DNA microarray classification. Neurocomputing 2006; 70: 187 – 99.
7. Bilge K, Rajeev R, Wesley E, Snyder. A comparative analysis of structural risk minimization by support vector machines and nearest neighbor rule. Pattern Recogn Lett 2004; 25: 63 – 71.
8. Cover TM. Estimation by the nearest neighbor rule. IEEE T., Inform. Theory, 14, (1968) 50 – 5.
9. Tae YS, Jeong WS, Kong MH, Lee JS, Park SB, Lee SJ. A hybrid approach to error reduction of support vector machines in document classification. Proceedings of the Third International Conference on Information Technology: New Generations (ITNG'06). 2006; 501 – 6.
10. Acevedo FJ, Maldonado DSE, Narváez A, López,F. Probabilistic support vector machines for multi-class alcohol identification. Sensor Actuat B-Chem 2007; 122: 227 – 35.
11. John ST, Nello C. Kernel Methods for Pattern Analysis. Cambridge University Press. 2004; 24 – 32.
12. Cheong S, Sang HO, Lee SY. Support vector machines with binary tree architecture for multi-class classification. Neural Inform Process – Lett and Rev 2004; 2: 47 – 51.
13. Richard O, Duda PE, Hart DG. Stork Pattern Classification. Second Edition. John Wiley and Sons, Inc. 2000; 131 – 58.
14. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson J, Marks J, Nevins J. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci USA 2001; 98: 11462 – 7.
15. Spang R, Zuzan H, West M, Nevins J, Blanchette C, Marks JR. Prediction and uncertainty in the analysis of gene expression profiles. In Silic Biol 2002; 2: 369 – 81.
16. Yeoh E, Mary ER, Sheila AS, Williams WK. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell 2002; 1: 133 – 43.