International Journal of Neural Systems, Vol. 15, No. 6 (2005) 475–484

World Scientific
www.worldscientific.com

APPLICATIONS OF SUPPORT VECTOR MACHINES TO CANCER CLASSIFICATION WITH MICROARRAY DATA

FENG CHU and LIPO WANG
School of Electrical and Electronic Engineering,
Nanyang Technological University,
Block S1, Nanyang Avenue, Singapore 639798
elpwang@ntu.edu.sg

Microarray gene expression data usually have a large number of dimensions, e.g., over ten thousand genes, and a small number of samples, e.g., a few tens of patients. In this paper, we use the support vector machine (SVM) for cancer classification with microarray data. Dimensionality reduction methods, such as principal components analysis (PCA), class-separability measure, Fisher ratio, and t-test, are used for gene selection. A voting scheme is then employed to do multi-group classification by $k(k-1)$ binary SVMs. We are able to obtain the same classification accuracy but with much fewer features compared to other published results.

Keywords: Cancer classification; gene expression data; microarray; support vector machine.

1. Introduction

Microarrays,[16] also known as gene chips or DNA chips, provide a convenient way of obtaining gene expression levels for a large number of genes simultaneously. Each spot on a microarray chip contains the clone of a gene from a tissue sample. Some mRNA samples are labelled with two different kinds of dyes, for example, Cy5 (red) and Cy3 (blue). After mRNA interact with the genes, i.e., hybridization, the color of each spot on the chip will change. The resulted image reflects the characteristics of the tissue at the molecular level.

Microarrays can thus be used to help classify and predict different types of cancers. Traditional methods for diagnosis of cancers are mainly based on the morphological appearances of the cancers; however, sometimes it is extremely difficult to find clear distinctions between some types of cancers according to their appearances. Hence the microarray technology stands to provide a more quantitative means for cancer diagnosis. For example, gene expression data have been used to obtain good results in the classifications of lymphoma,[1] leukemia,[10] breast cancer,[14] and liver cancer.[4]

It is challenging to use gene expression data for cancer classification because of the following two special aspects of gene expression data. First, gene expression data are usually very high dimensional. The dimensionality ranges from several thousands to over ten thousands. Second, gene expression data sets usually contain relatively small numbers of samples, e.g., a few tens. If we treat this pattern recognition problem with supervised machine learning approaches, we need to deal with the shortage of training samples and high dimensional input features. Recent approaches to solve this problem include artificial neural networks,[13] an evolutionary algorithm,[7] nearest shrunken centroids,[18] and a graphical method.[3]

In this paper, we apply a powerful classifier, i.e., the support vector machine (SVM), and four effective feature reduction methods, i.e., principal components analysis (PCA), class-separability measure, Fisher ratio, and t-test, to the problem of cancer classification based on gene expression data. This paper is organized as follows. Three gene expression data sets used in this paper are described in Sec. 2. Then we discuss an important step for dimension reduction, i.e., gene selection, in Sec. 3. Numerical results

are presented in Sec. 4, followed by discussions and conclusions in Sec. 5.

2. Gene Expression Data Sets

2.1. *The SRBCT data set*

The SRBCT data set[13] can be obtained from the website (http://research.nhgri.nih.gov/microarray/Supplement/). The entire data set includes the expression data of 2308 genes. There are totally 63 training samples and 25 testing samples, five of the testing samples being not SRBCTs. The 63 training samples contain 23 Ewing family of tumors (EWS), 20 rhabdomyosarcoma (RMS), 12 neuroblastoma (NB), and 8 Burkitt lymphomas (BL). And the 20 SRBCTs testing samples contain 6 EWS, 5 RMS, 6 NB, and 3 BL.

2.2. *The lymphoma data set*

The lymphoma data set[1] can be obtained from the website (http://llmpp.nih.gov/lymphoma). In this data set, there are 42 samples derived from diffuse large B-cell lymphoma (DLBCL), 9 samples from follicular lymphoma (FL), and 11 samples from chronic lymphocytic lymphoma (CLL). The entire data set includes the expression data of 4026 genes. In this data set, a small part of data is missing. A k-nearest neighbor algorithm was applied to fill those missing values.[20]

2.3. *The leukemia data set*

The leukemia data set[10] can be obtained at (http://www-genome.wi.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43). The samples in this data set belong to two types of leukemia, i.e., the acute myeloid leukemia (AML) and the acute lymphoblastic leukemia (ALL). Among these samples, 38 of them are used for training and the other 34 independent samples are for testing. The entire leukemia data set contains the expression data of 7129 genes.

Ordinarily, raw gene expression data should be normalized to reduce the systemic bias introduced during experiments. For the SRBCT and the lymphoma data sets, the data after normalization can be found on the web. However, for the leukemia data set, normalized data are not available. Thereafter, we need to do normalization by ourselves.

We followed the normalization procedure used in (Ref. 8). Three steps were taken, i.e., (a) thresholding: with a floor of 100 and a ceiling of 16000, that is, if a value is greater/smaller than the ceiling/floor, this value is replaced by the ceiling/floor; (b) filtering, leaving out the genes with $max/min \leq 5$ or $(max - min) \leq 500$, here max and min refer to the maximum and minimum of the expression values of a gene, respectively; (c) carrying out logarithmic transformation with 10 as the base to all the expression values. 3571 genes survived after these three steps. Furthermore, the data were standardized across experiments, i.e., minus the mean and divided by the standard deviation of each experiment.

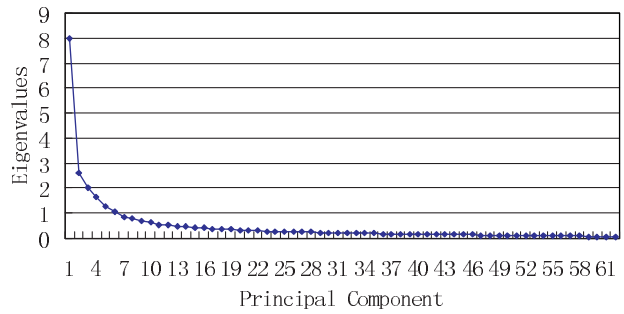
3. Gene Selection Methods

3.1. *Introduction*

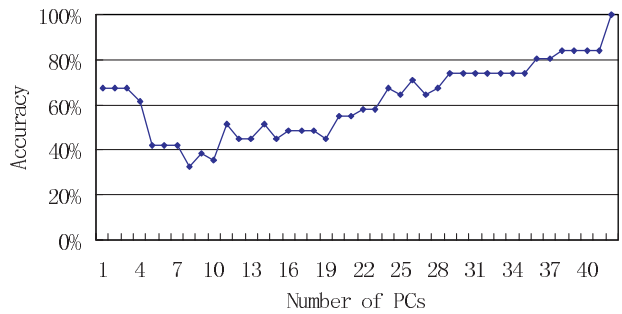
Among the large number of genes, only a small part may benefit the correct classification of cancers. The rest of the genes have little impact on the classification. Even worse, some genes may act as "noise" and undermine the classification accuracy. Hence, to obtain good classification accuracy, we need to pick out the genes that benefit the classification most. In addition, gene selection is also a procedure of input dimension reduction, which leads to a much less computation load to the classifier. Maybe more importantly, reducing the number of genes used for classification can help researchers put more attention on these important genes and find the relationship between those genes and the development of the cancers.

3.2. *Principal component analysis*

The most widely used technique for input dimension reduction in gene expression analysis is principal component analysis (PCA) (see, e.g., Chapter 8 of (Ref. 17)). The basic idea of principal component analysis is transforming the input space into a new space described by the principal components (PCs). All the PCs are orthogonal to each other and are ordered according to the absolute values of their eigenvalues. The kth PC is the vector with the kth largest eigenvalue. By leaving out the vectors with small eigenvalues, the dimensionality of the input space is reduced. Because PCA chooses vectors with the largest eigenvalues, it covers the directions with the largest vector variations in the input space.



(a)



(b)

Fig. 1. Principal components analysis for the lymphoma data set. (a): the change of the eigenvalues; (b): the classification result using PCA for input reduction. The horizontal axis is the number of the PCs with the greatest eignvalues that are used.

In the directions determined by the vectors with small eigenvalues, the vector variations are very small. In a word, PCA intends to capture the most informative directions. Figure 1a shows the change of eigenvalues in the lymphoma data set. The classification result using PCA as input dimension reduction scheme is given in Fig. 1b. Through comparing Fig. 1b with Fig. 8a, we found that t-test (to be introduced in the latter part) could achieve much better classification accuracy than PCA.

3.3. *Class-separability analysis for gene selection*

Another frequently used method for gene selection is to measure the class-separability (CS).[8] CS of gene i is defined as:

$$CS_i = SB_i/SW_i, \tag{1}$$

where

$$SB_i = \sum_{i=1}^{K} (\bar{x}_{ik} - \bar{x}_i)^2, \tag{2}$$

$$SW_i = \sum_{k=1}^{K} \sum_{j \in C_k} (x_{ij} - \bar{x}_{ik})^2, \tag{3}$$

$$\bar{x}_{ik} = \sum_{j \in C_k} \bar{x}_{ij}/n_k \tag{4}$$

$$\bar{x}_i = \sum_{j=1}^{n} x_{ij}/n. \tag{5}$$

Here SB_i is the sum of squares of between-class distances (the distances between samples of different classes). SW_i is the sum of squares of with-in class distances (the distances of samples within the same class). In the whole data set, there are K classes. C_k refers to class k that includes n_k samples. x_{ij} is the expression value of gene i in sample j. $\bar{x}_{ik}$ is the mean expression value in class k for gene i. n is the total number of samples. $\bar{x}_i$ is the general mean expression value for gene i. A CS is calculated for each gene. A larger CS indicates a larger ratio of the distances between different classes to the distances within one specific class. Therefore, CSs can be used to measure the capability of genes to separate different classes. 20 important genes with the largest CSs in the SRBCT data set are given in Table 1.

3.4. *Fisher-ratio for gene selection*

Fisher ratio (see e.g., (Ref. 15)) is also a ratio of between-class distances to with-in class distances. If there are two classes in a data set, the Fisher ratio (F) for gene i is:

$$F_i = \frac{(\bar{x}_{i1} - \bar{x}_{i2})^2}{{s_{i1}}^2 + {s_{i2}}^2}. \tag{6}$$

If there are more than two classes in the data set,

$$F_i = \sum_{p=1}^{K} \sum_{q=1}^{K} \frac{(\bar{x}_{ip} - \bar{x}_{iq})^2}{K(K-1)({s_{ip}}^2 + {s_{iq}}^2)}, \quad p \neq q. \tag{7}$$

where

$$s_{ik}^2 = \frac{1}{n_k} \sum_{j \in C_k} (x_{ij} - \bar{x}_{ik})^2 \tag{8}$$

The main differences between the CS and the F are as follows. (a) When extending F from two-class version in Eq. (6) to multi-class version, Eq. (7) averages all the F between any two different classes; however, CS can be directly calculated for multi-class data sets. (b) When calculating the between-class distance, the CS adds up all the distances from

Table 1. The 20 top genes selected by class separability in the SRBCT data set.

Rank	Gene ID	Gene description
1	770394	Fc fragment of IgG, receptor, transporter, alpha
2	796258	sarcoglycan, alpha (50 kD dystrophin-associated glycoprotein)
3	784224	fibroblast growth factor receptor 4
4	814260	follicular lymphoma variant translocation 1
5	295985	ESTs
6	377461	caveolin 1, caveolae protein, 22 kD
7	859359	quinone oxidoreductase homolog
8	769716	neurofibromin 2 (bilateral acoustic neuroma)
9	365826	growth arrest
10	1435862	antigen identified by monoclonal antibodies 12E7, F21 and O13
11	866702	protein tyrosine phosphatase, non-receptor type 13 (APO-1/CD95 (Fas)-associated phosphatase)
12	296448	insulin-like growth factor 2 (somatomedin A)
13	740604	interferon stimulated gene (20 kD)
14	241412	E74-like factor 1 (ets domain transcription factor)
15	810057	cold shock domain protein A
16	244618	ESTs
17	52076	olfactomedinrelated ER localized protein
18	21652	catenin (cadherin-associated protein), alpha 1 (102kD)
19	43733	glycogenin 2
20	236282	Wiskott-Aldrich syndrome (ecezema-thrombocytopenia)

the mean of each class ($\bar{x}_{ik}$) to the overall mean ($\bar{x}_i$); however, F directly calculates the distances between the means of any two different classes and then adds them up.

3.5. *A t-test-based gene selection approach*

T-test is a statistical method proposed by Welch.[23] It is used to measure how large the difference is between the distributions of two groups of samples. For a specific gene, if it shows larger distinctions between 2 groups, it is more important for the classification of the two groups. To find the genes that contribute most to the classification, t-test has been used in gene selection in recent years.[21]

To select important genes using t-test involves several steps. In the first step, a score based on t-test (named t-score or TS) is calculated for each gene. In the second step, all the genes are rearranged according to their TSs. The gene with the largest TS is put in the first place of the ranking list, followed by the gene with the second largest TS, and so on. Finally, only some top genes in the list are used for classification.

The standard t-test is only applicable to measure the difference between two groups. Therefore, when the number of classes is more than two, we need to modify the standard t-test. In this case, we use t-test to calculate the degree of difference between one specific class and the centroid of all the classes. Hence, the definition of TS for gene i can be described like this:

$$TS_i = max\left\{\left|\frac{\bar{x}_{ik} - \bar{x}_i}{m_k s_i}\right|, k = 1, 2, \ldots, K\right\} \quad (9)$$

$$s_i^2 = \frac{1}{n-K}\sum_k \sum_{j \in C_k} (x_{ij} - \bar{x}_{ik})^2 \quad (10)$$

$$m_k = \sqrt{1/n_k + 1/n} \quad (11)$$

Here $max\{y_k, k = 1, 2, \ldots, K\}$ is the maximum of all $\{y_k, k = 1, 2, \ldots, K\}$. s_i is the pooled within-class standard deviation for gene i.

The classification results using the 20 top genes selected by class-separability, Fisher-ratio, and the t-test in the SRBCT data set are shown in Fig. 2. From these results, it is obvious that t-test is better than CS and Fisher-ratio for gene selection.

4. Experimental Results

The structure of an SVM[2,5,22] is shown in Fig. 3. Two of the most commonly used kernel functions are the

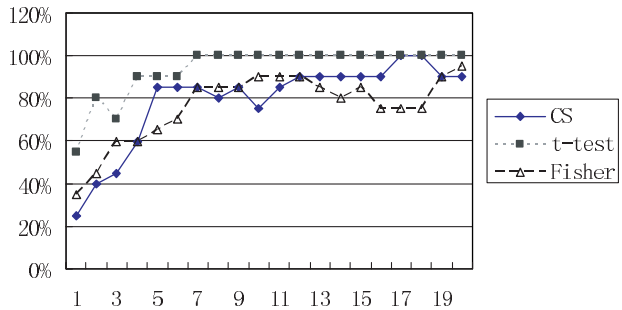


Fig. 2. Comparison of classification results in the SRBCT data set using the 20 genes selected by the class-separability, the t-test, and the Fisher ratio approaches.

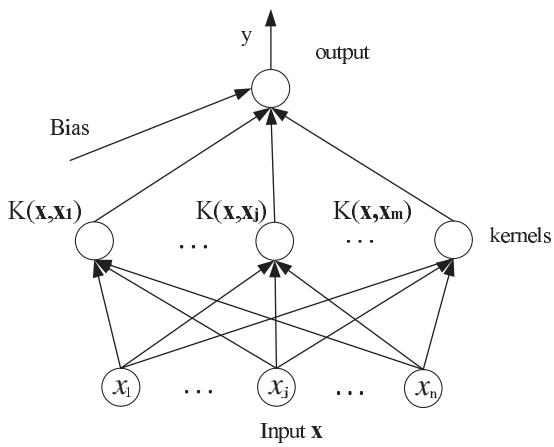


Fig. 3. The structure of an SVM.

polynomial kernel:

$$K(\mathbf{x}, \mathbf{x}_i) = (\mathbf{x}^T \mathbf{x}_i + 1)^p \tag{12}$$

where p is a constant specified by users, and the radial basis function:

$$K(\mathbf{x}, \mathbf{x}_i) = exp(-\gamma \|\mathbf{x} - \mathbf{x}_i\|^2) \tag{13}$$

where γ is also a constant specified by users.

If there are more than two classes in the data set, binary SVMs are not sufficient to solve the whole problem. To solve multi-class classification problems, we should divide a whole problem into a number of binary classification problems. Usually, there are two approaches.[12] One is the "one against all" scheme and the other is the "one against one" scheme.

For the "one against all" scheme, if there are N classes in the entire data set, then N independent binary classifiers are built. Each binary classifier is in charge of picking out one specific class from all the other classes. For one specific pattern, all the N classifiers are used to make a prediction. The pattern is categorized to the class that receives the strongest prediction. The prediction strength is measured by the result of the decision function.

For the "one against one" scheme, there must be one (and only one) classifier taking charge of the classification between any two classes. Therefore, for a data set with K classes, $K(K-1)/2$ binary classifiers are used. To get the ultimate result, a voting scheme is used. For every input vector, all the classifiers give their votes so there will be $K(K-1)/2$ votes, when all the classification (voting) finished, the vector is designated to the class getting the highest number of votes. If a vector gets highest votes for more than one class, it is randomly designated to one of them. In fact, there is still no conclusion about which scheme is better for combining binary classifiers.[11] In our practice, we choose the "one against one" scheme.

We applied the SVM described above to process the SRBCT, the lymphoma, and the leukemia data sets. The results are as follows.

4.1. *Results for the SRBCT data set*

In this data set, we first ranked the importance of all the genes with TSs. We picked out 60 of them with the largest TSs to do classification. The top 30 genes are listed in Table 2. We input these genes one by one to the SVM classifier according to their ranks. That is, we first input the gene ranked No.1 in Table 2. Then, we trained the SVM classifier with the training data and tested the SVM classifier with the testing data. After that, we repeated the whole process with top 2 genes, and then top 3 genes, and so on. Figure 4 shows the training and the testing accuracies with respect to the number of genes used.

In this data set, we used SVMs with RBF kernels. C and γ were set as 80 and 0.005, respectively. This classifier obtained 100% training accuracy and 100% testing accuracy using the top 7 genes. Actually, the values of C and γ have great impact on the classification accuracy. Figure 5 shows the classification results with different values of γ(gamma). We also applied linear SVMs and SVMs with polynomial kernel function to the SRBCT data set. The results are shown in Fig. 6 and Fig. 7. The linear SVMs and the SVMs with the polynomial kernel function obtained 100% accuracy with 7 and 6 genes, respectively. The similarity of these results indicates that

Table 2. The 30 top genes selected by t-test in the SRBCT data set.

Rank	Gene ID	Gene description
1	810057	cold shock domain protein A
2	784224	fibroblast growth factor receptor 4
3	296448	insulin-like growth factor 2 (somatomedin A)
4	770394	Fc fragment of IgG, receptor, transporter, alpha
5	207274	Human DNA for insulin-like growth factor II (IGF-2); exon 7 and additional ORF
6	244618	ESTs
7	234468	ESTs
8	325182	cadherin 2, N-cadherin (neuronal)
9	212542	Homo sapiens mRNA; cDNA DKFZp586J2118 (from clone DKFZp586J2118)
10	377461	caveolin 1, caveolae protein, 22kD
11	41591	meningioma (disrupted in balanced translocation) 1
12	898073	transmembrane protein
13	796258	sarcoglycan, alpha (50 kD dystrophin-associated glycoprotein)
14	204545	ESTs
15	563673	antiquitin 1
16	44563	growth associated protein 43
17	866702	protein tyrosine phosphatase, non-receptor type 13 (APO-1/CD95 (Fas)-associated phosphatase)
18	21652	catenin (cadherin-associated protein), alpha 1 (102 kD)
19	814260	follicular lymphoma variant translocation 1
20	298062	troponin T2, cardiac
21	629896	microtubule-associated protein 1B
22	43733	glycogenin 2
23	504791	glutathione S-transferase A4
24	365826	growth arrest-specific 1
25	1409509	troponin T1, skeletal, slow
26	1456900	Nil
27	1435003	tumor necrosis factor, alpha-induced protein 6
28	308231	Homo sapiens incomplete cDNA for a mutated allele of a myosin class I, myh-1c
29	241412	E74-like factor 1 (ets domain transcription factor)
30	1435862	antigen identified by monoclonal antibodies 12E7, F21 and O13

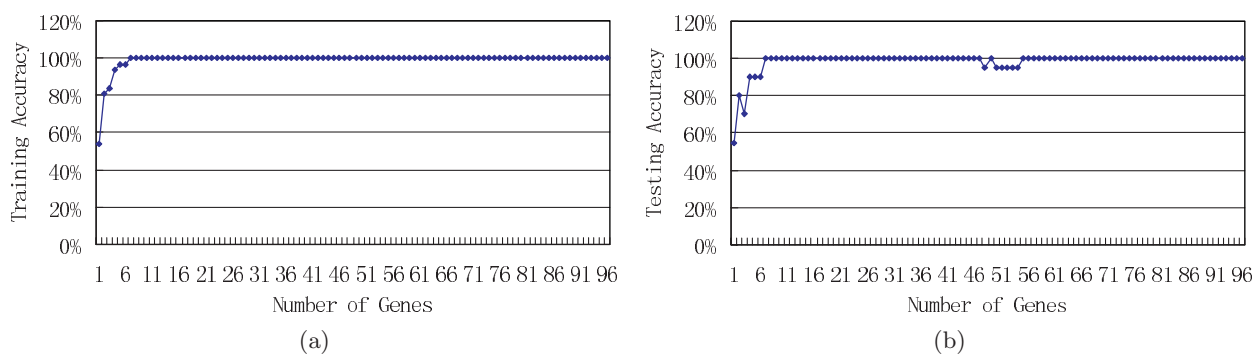


Fig. 4. The classification results for the SRBCT data set: (a) the training accuracy; (b) the testing accuracy.

the SRBCT data set is separable for all the three kinds of classifiers.

For the SRBCT data set, Khan *et al.*[13] 100% accurately classified the 4 types of cancers with a linear artificial neural network by using 96 genes. Their results and our results of the linear SVMs both proved that the classes in the SRBCT data set are linearly separable. In 2002, Tibshirani *et al.*[18]

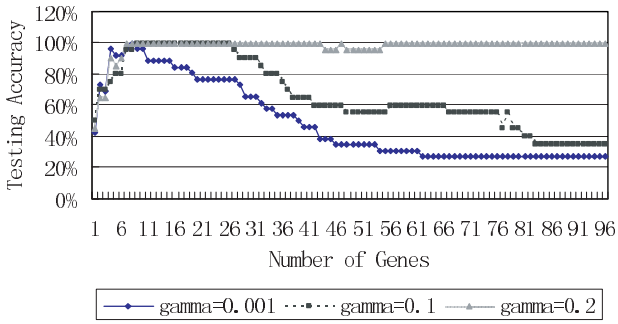


Fig. 5. The testing results of SVMs with the RBF kernels in different values of γ.

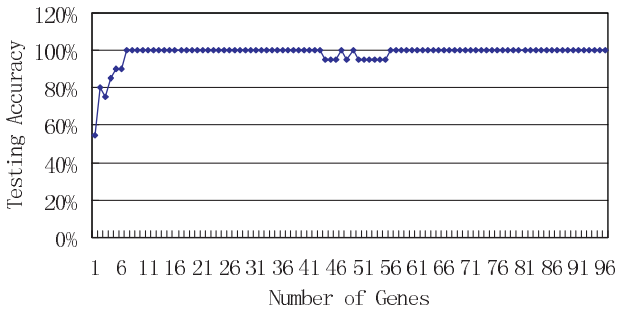


Fig. 6. The testing results of the linear SVMs for the SRBCT data set.

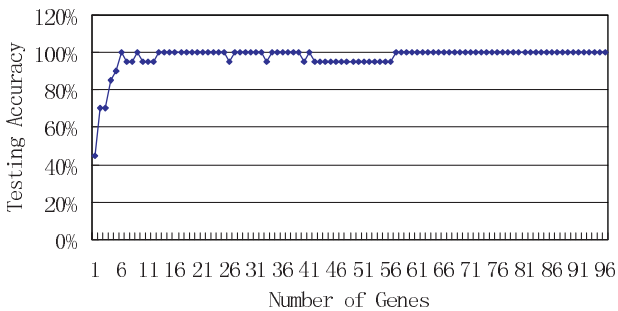


Fig. 7. The testing result of the SVMs with the polynomial kernel function ($p = 2$) for the SRBCT data set.

also correctly classified the SRBCT data set with 43 genes by using a method named nearest shrunken centroids. Deutsh[7] further reduced the number of genes needed for reliable classification to 12 with an evolutionary algorithm. Compared with these previous results, the SVMs introduced here can achieve 100% accuracy with 7 genes (for the linear and the RBF kernel function version) and 6 genes (for the polynomial kernel function version). Table 3 gives the detail of this comparison.

Table 3. Comparison of number of genes required by different methods to achieve 100% classification accuracy.

Method	Number of genes
MLP ANN (Khan)	96
Nearest Shrunken Centroids	43
Evolutionary Algorithm	12
SVM (linear or RBF kernel function)	7
SVM (polynomial kernel function, $p = 2$)	6

4.2. *Results for the lymphoma data set*

In the lymphoma data set, the top 196 genes selected by TSs are listed in Table 4. Figures 8a and 8b are training and testing results of the top 70 genes. The classifiers used here are also RBF SVMs. C and γ are 20 and 0.1, respectively. They obtained 100% accuracy in both training and testing data set with only 5 genes.

In the lymphoma data set, nearest shrunken centroids[19] used 48 genes to give a 100% accurate classification. In comparison with this, the SVMs we used greatly reduced the number of genes required.

4.3. *Results in the leukemia data set*

The leukemia data set is a widely known data set. Golub *et al.*, built a 50-gene classifier[10] for it. This classifier made 1 error in the 34 testing samples; and in addition, it cannot give strong prediction to the other 3 samples. Nearest shrunken centroids made 2 errors among the 34 testing samples. From Fig. 9a and Fig. 9b, we found the RBF SVMs we used also made 2 errors in the testing data set with 20 genes.

5. Conclusions

To find a good solution to the problem of cancer classification using gene expression data, one can work towards two related directions. One is gene selection. Selecting important genes can make the task easier because important genes determine a new input space in which the samples are more likely to be correctly classified. The other direction is to build powerful classifiers.

Table 4. The top 196 important genes selected by *t*-test for the lymphoma data set.

Rank	Gene ID	Gene description
1	GENE1610X	Mig=Humig=chemokine targeting T cells
2	GENE708X	Ki67 (long type)
3	GENE1622X	CD63 antigen (melanoma 1 antigen)
4	GENE1641X	Fibronectin 1
5	GENE3320X	Similar to HuEMAP=homolog of echinoderm microtubule associated protein EMAP
6	GENE707X	Topoisomerase II alpha (170 kD)
7	GENE653X	Lactate dehydrogenase A
8	GENE1636X	Fibronectin 1
9	GENE2391X	Unknown
10	GENE2403X	Unknown
11	GENE1644X	cathepsin L
12	GENE3275X	Unknown UG Hs.192270 ESTs
13	GENE642X	nm23-H1=NDP kinase A=Nucleoside dephophate kinase A
14	GENE706X	CDC2=Cell division control protein 2 homolog=P34 protein kinase
15	GENE1643X	cathepsin L
16	GENE2395X	Unknown UG Hs.59368 ESTs
17	GENE537X	B-actin,1099-1372
18	GENE709X	STAT induced STAT inhibitor-1=JAB=SOCS-1
19	GENE2307X	CD23A=low affinity II receptor for Fc fragment of IgE
20	GENE2389X	Unknown
...	...	...
...	...	...
187	GENE646X	nm23-H2=NDP kinase B=Nucleoside dephophate kinase B
188	GENE2180X	Unknown
189	GENE506X	putative oral tumor suppressor protein (doc-1)
190	GENE632X	ATP5A=mitochondrial ATPase coupling factor 6 subunit
191	GENE844X	ets-2=ets family transcription factor
192	GENE629X	HPRT=IMP:pyrophosphate phosphoribosyltransferase E.C. 2.4.2.8.
193	GENE2381X	Arachidonate 5-lipoxygenase=5-lipoxygenase=5-LO
194	GENE1533X	CD11C=leukocyte adhesion protein p150,95 alpha subunit=integrin alpha-X
195	GENE2187X	SF1=splicing factor
196	GENE641X	cell cycle protein p38-2G4 homolog (hG4-1)

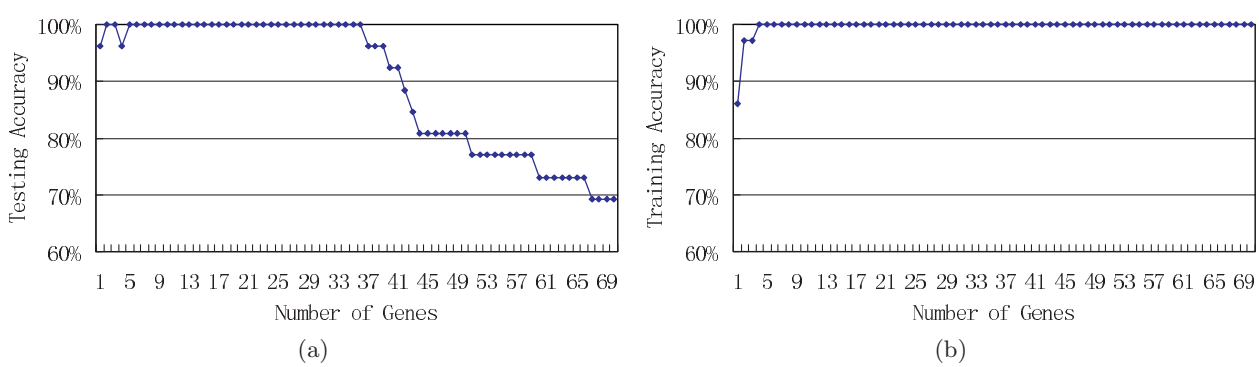
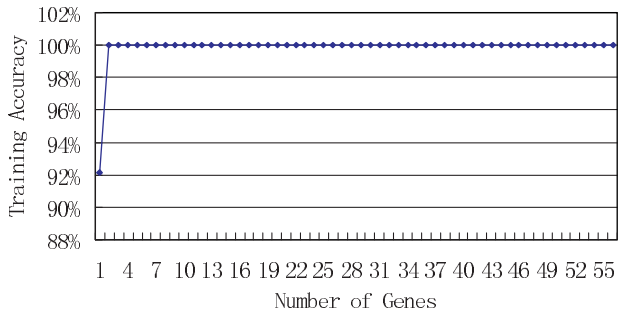
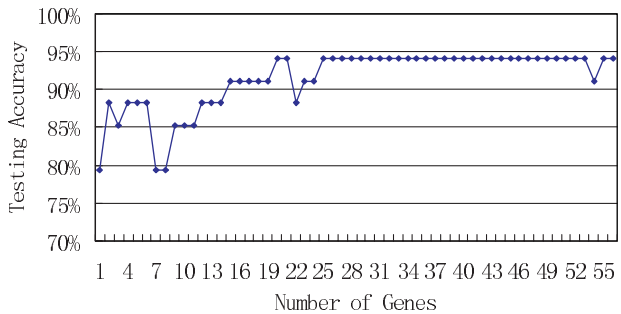


Fig. 8. The classification results for the lymphoma data set: (a) the testing accuracy; (b) the training accuracy.

In this paper, we touched upon both directions from an experimental viewpoint. For gene selection, we tested 4 well known schemes, i.e., PCA, class separability analysis, Fisher ratio, and *t*-test. Our results showed that *t*-test-based gene selection outperformed the other three approaches. Therefore, we applied this method to select important genes.



(a)



(b)

Fig. 9. The classification results for the leukemia data set: (a) the training accuracy; (b) the testing accuracy.

This paper also described the procedure of constructing support vector machines. After that, 3 gene expression data sets were classified with the SVMs we built. The results proved the effectiveness of the SVMs. In all the three data sets, the SVMs obtained very high classification accuracies with much fewer genes compared with previously published methods. In addition, we found that effective gene selection schemes could largely simplify the requirements to classifiers. For example, linear SVMs achieved high accuracy in the SRBCT data set. However, this does not mean the SVM can not deal with nonlinear problems. In fact, SVMs with RBF or polynomial kernels are good classifiers for nonlinear problems (see for e.g., Ref. 22).

References

1. A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma and I. S. Lossos *et al.*, Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling, *Nature* **403** (2000) 503–511.
2. B. Boser, I. Guyon and V. N. Vapnik, A training algorithm for optimal margin classifiers, in *Fifth Annual Workshop on Computational Learning Theory*, (Pittsburgh, USA), (ACM Press), (1992) 144–152.
3. E. Bura and R. M. Pfeiffer, Graphical methods for class prediction using dimension reduction techniques on DNA microarray data, *Bioinformatics* **19** (2003) 1252–1258.
4. X. Chen, S. T. Cheung, S. So, S. T. Fan and C. Barry, Gene expression patterns in human liver cancers, *Molecular Biology of Cell* **13** (2002) 1929–1939.
5. C. Cortes and V. Vapnik, Support vector networks, *Machine Learning* **20** (1995) 273–297.
6. T. M. Cover, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, *IEEE Trans. Electronic Computers EC* **14** (1965) 326–334.
7. J. M. Deutsch, Evolutionary algorithms for finding optimal gene sets in microarray prediction, *Bioinformatics* **19** (2003) 45–52.
8. S. Dudoit, J. Fridlyand and T. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, *J. Am. Stat. Assoc.* **97** (2002) 77–87.
9. R. N. Fletcher, Practical Methods of Optimization, 2nd edn. (Wiley, New York, 1987).
10. T. Golub, D. K. Slonim, P. Tamayo, C. Huard and M. Gaasenbeek *et al.*, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* **286** (1999) 531–536.
11. H. J. Hastie and R. J. Tibshirani, Classification by pairwise coupling, in *Advances in Neural Information Processing Systems*, eds. M. I. Jordan, M. J. Kearnsa and S. A. Solla, Vol. 10 (MIT Press, 1998).
12. S. Knerr, L. Personnaz and G. Dreyfus, Single layer learning revisited: A stepwise procedure for building and training neural network, in *Neurocomputing: Algorithms, Architectures and Applications*, ed. J. Fogelman (Springer-Verlag, 1990).
13. J. Khan, J. S. Wei, M. Ringner, L. H. Saal and M. Ladanyi *et al.*, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nature Medicine* **7** (2001) 673–679.
14. X. J. Ma, R. Salunga, J. T. Tuggle, J. Gaudet and E. Enright *et al.*, Gene expression profiles of human breast cancer progression, in *Proc. Natl. Acad. Sci. USA*, Vol. 100 (2003), pp. 5974–5979.
15. K. Z. Mao, RBF neural network center selection based on Fisher ratio class separability measure, *IEEE Trans. on Neural Networks* **13** (2002) 1211–1217.
16. M. Schena, D. Shalon, R. W. Davis and P. O. Brown, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science* **267** (1995) 467–470.
17. H. Simon, Neural Networks: A Comprehensive Foundation, 2nd edn. (Prentice-Hall Inc., New Jersey, 1999).
18. R. Tibshirani, T. Hastie, B. Narashiman and G. Chu, Diagnosis of multiple cancer types by shrunken

centroids of gene expression, in *Proc. Natl. Acad. Sci. USA*, Vol. 99 (2002), pp. 6567–6572.
19. R. Tibshirani, T. Hastie, B. Narasimhan and G. Chu, Class predicition by nearest shrunken centroids with applications to DNA microarrays, *Statistical Science* **18** (2003) 104–117.
20. O. Troyanskaya, M. Cantor, G. Sherlock *et al.*, Missing value estimation methods for DNA microarrays, *Bioinformatics* **17** (2001) 520–525.
21. V. G. Tusher, R. Tibshirani and G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, in *Proc. Natl. Acad. Sci. USA* **98** (2001) 5116–5121.
22. V. N. Vapnik, Statistical Learning Theory (Wiley, New York, 1998).
23. B. L. Welch, The generalization of student's problem when several different population are involved, *Biomethika* **34** (1947) 28–35.