Minimizing Essential Set Based Feature Selection for Cancer Classification

Khin May Win, Nan Sai Moon Kham
University of Computer Studies, Yangon
winn.km05@gmail.com, moonkhamucsy@gmail.com

Abstract

Many methods for classification and gene selection with microarray data have been developed. Some methods usually give a ranking of genes. Relevant gene rank criteria is derived from SVM and based on generalization error bounds with respect to genes variable .We address feature selection problem for classification because of small samples with high dimensionality of genes. The best choice of gene subset means selection of relevant features that is a key for building a more accurate classifier. We propose a new method using minimizing essential set (MES) generated based on the nearest neighbor rule. It is related to structural risk minimization and thus leads to good generalization. The proposed method is compared to some standard feature selection methods with three real datasets. Our approach is computationally efficient with better classification performance.

Keywords: *DNA microarray, feature selection, cancer classification ,recursive feature elimination, support vector machines.*

1. Introduction

In recent years, the rapid development of DNA Microarray technology has made it possible for scientist to monitor the expression level of thousands of genes. Especially accurate classification of cancer disease is very important for treatment of cancer. Many researchers have been studying many problems of cancer classification using gene expression data to prove the optimal classification. Support Vector Machines (SVMs) and related kernel methods have become increasingly popular tools for data mining tasks such as classification and regression. But the limitation is overfitting problems in classification because of the huge microarray data. As one of the most commonly used learning methods, SVM has shown excellent performance in handling the high dimensional feature space [2]. Feature selection is a fairly straightforward procedure for linear or nonlinear support vector machine (SVM) classifiers. For example, a 1-norm support vector machine linear classifier obtained by concave minimization will easily reduce the features [4]. Selecting genes that are informative for the classification is one key issue for understanding the biology behind the classification and an important step towards discovering those genes are responsible for the cancer identification .We should minimize generalization error on the expected risk by controlling both the training errors and the capacity of the set of prospect functions measured by the so-called Vapnik–Chervonnkis dimensions [15]. In this paper, using the primary concepts behind the SVM approach by examination the dimensional reduction problem of classification. There has been considerable recent interest in feature selection for SVMs. Weston has proposed reducing features based on minimizing generalization bounds via a gradient descent approach [14]. H. Stoppiglia and G. Dreyfus introduced an incremental approach based on ranking features by their effect on the margin of hyperplane [7]. Another approach based on a Bayesian interpretation of SVMs is presented by Goldet al [9]. For instance, Weston et al. proposed a method based on finding the best variable subset which minimizes the margin bounds [13]. One approach based on smoothing spleen ANOVA kernels is proposed by Zhang [17]. Guyon use a wrapper method designed for SVMs [6]. Another possibility is to use a filter method such as Relief in conjunction with an SVMs [5].A representative method of this approach is recursive feature elimination (RFE) based on support vector machines (SVMs) aspect, which uses linear SVM to classify the samples and ranks the genes in the classifier by their weights. The purpose of feature selection is to eliminate irrelevant variables to enhance the generalization performance and how SVMs can perform badly in the situation of many irrelevant features. Feature selection is the process of searching for a subset of relevant features from a larger set of original ones preferring classification performance or class separability. In fact, feature selection methods play a significant role for solving cancer sample classification problems where the number of features is much larger than the number of samples.

To consider possible combinatorial effects of genes, most wrapper methods adopt more sophisticated multivariate machine learning strategies such as SVMs and neural networks [7][10].These methods have been shown in many experiments to be more powerful in terms of classification accuracy. Evaluating the statistical significance of the detected feature is the central idea in the paradigm of statistical inference from experimental data [6]. For the ranking of gene, it is observed by a certain method if the gene is informative or not informative to the classification. If informative, that is in the sense of the criteria defined or implied by the

classification and ranking method. We call this problem the significance of gene ranking or feature ranking. We raise this problem in this paper and describe our strategy towards a solution for microarray classification of cancer samples.

We propose the strong feature selection property by the new algorithm with SVMs. After formally stating the problem and reviewing the prior work related with feature selection and SVMs in (section 2), a new method is presented at (section 3) by gene selection using SVMs. Then we study the data set and provide the basic of our experimental result at (section 4). Finally we conclude our work and tend to future work at (section 5).

2. The Feature Selection Problem

In this paper, the problem is to find way to reduce the dimensionality of the feature space. Data overfitting arises when the number of feature is very large. To overcome this problem, projecting method used at Latent models is the first new method used to reduce feature space dimensionality [11].With such method, one disadvantage is none of the original input features can be discarded. In the literature review, one distinguishes between two types of method to solve the generalization problem; so called filter and wrapper methods. Filter methods are defined as preprocessing steps to induction that can remove irrelevant attributes before induction occur, to be valid for any set of gene data. For example one popular filter method is to use Pearson correlation coefficients [2].

Another one is wrapper method , is defined as a search through the space of feature subset using the estimated accuracy from an induction algorithm as a measure of goodness of a particular feature subset. Thus, the way to approximates the error is function $f(\sigma,\alpha)$ by minimizing ,

$$T_{wrap}(\sigma,\alpha) = \min_{\sigma} T_{\text{alg}}(\sigma) \qquad (1)$$

subject to $\sigma \in \{0,1\}^n$ where T_{alg} is a learning algorithm trained on data preprocessed with fixed σ [10] .Obviously, wrapper method can provide more accurate solutions than filter method but in general are more computationally expensive.

For some classification problems, the ideal objective function is the expected value of the error , that is the error rate computed on an infinite number of example samples. The idea is to compute the change in function f caused by removing a given feature.

The feature selection problem can be addressed in the following two ways (1) given a fixed m<<n, where n is the original number of features, find the m features that give the smallest expected generalization error;(2) if given a maximum allowable generalization error σ ,to find the smallest m. This problem is formulated as follows.

Given a fixed set of function $y = f(x, \alpha)$ we wish to find a preprocessing of the data $x \rightarrow (x * \sigma)$ and the parameters α of the function f that give the minimum value of

$$ERR(\sigma,\alpha) = \int V(y, f((x*\sigma),\alpha))dP(x,y) \qquad (2)$$

subject to $\|\sigma\|_0 = m$, where P(x,y) is unknown, $(x*\sigma) = (x_1\sigma_1,.....,x_n\sigma_n)$ denotes an pairwise product, $V(.,.)$ is loss functional and $||.||_0$ is the 0-norm.

In this article, we introduce a feature selection algorithm for SVMs that take advantage of the performance increase and avoiding the computational complexity. Some previous work on feature selection for SVMs does exit, however results have been limited to linear kernels or linear probabilistic models[1][4].Our approach can be applied to linear or nonlinear problems. A better generalization can be achieved by replacing the smallest essential set to training data area.

3. Feature Selection with SVM Classification

Support vector machine (SVM) estimates the function classifying the data into two classes by Vapnik Theory [15]. SVM builds up a hyperplane as the decision surface in such a way to maximize the margin of separation between positive and negative examples. SVM achieves this by the structural risk minimization that is error rate of learning method on bounded by the sum of training error rate.

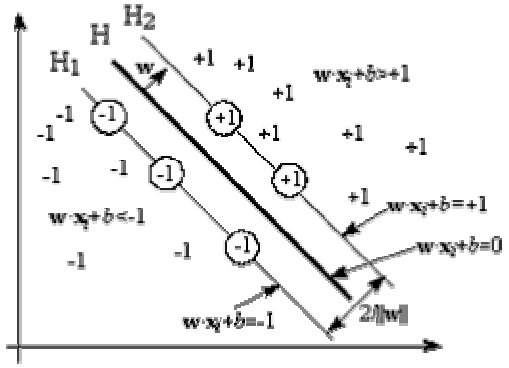

Figure.1. Optimum separation hyperplane of binary SVM

In this approach, support vector machine classifier is a binary classifier and linear separable that looks for an optimal hyperplane (H). For one training data set $\{x_k, y_k\} \in R^n \times \{-1,1\}$, where x_k are training examples and y_k the class labels. Then, computing a decision function of the form:

$$f(x) = (w, \Phi(x)) + b \quad (3)$$

The hyperplane parameterized by (w, b) while being consistent on the training set. The class label of **x** is obtained by considering the sign of *f(x)*.For the SVM classifier with misclassified examples being quadratically penalized, the optimization problem can be written as :

$$\min_{w,\xi} \frac{1}{2} \| w \|^2 + C \sum_{k=1}^{m} \xi_k^2 \quad (4)$$

under the constraint $\forall k$, $y_k f(x_k) \geq 1 - \xi_k$ .The solution of this problem is obtained using the Langrangian theory and the weight vector is of the form:

$$w = \sum_{k=1}^{m} \alpha_k y_k \Phi(x_k) \quad (5)$$

where α_k is the solution of the following quadratic optimization problem,

$$\max_{\alpha} W(\alpha) = \sum_{k=1}^{m} \alpha_k - \frac{1}{2} \sum_{k,l}^{m} \alpha_k \alpha_l y_k y_l (K(x_k, x_l) + \frac{1}{C} \delta_{k,l})$$

$$subject \quad to : \sum_{k=1}^{m} \alpha_k y_k \quad (6)$$

with vector $\boldsymbol{w}, \alpha_k$ is the solution of this problem and

$(\Phi(x_k), \Phi(x_l))$ is the Gram matrix of the training examples.

The generalization performance of classification is bounds on the leave-one-out error L. It is known to be an unbiased estimator of the generalization trained on *(m-1)* samples. One of the common L error bounds for SVMs is the margin bound (with nonzero bias b):

$$L = 4R^2 \| w \|^2 \quad (7)$$

where R is the radius of the smallest sphere that contains all the mapped data $\Phi(x_k)$.

3. Recursive Feature Elimination with SVM

For a linearly separable problem, SVMs find a discriminant function,$g(x_i)=w.x_i+b$, where b is the bias term, $x_i \in R^n$ are samples, and y_i are corresponding class labels $y_i = \{\pm 1\}, i = 1,..., m.$ .The discriminant function satisfies following constraint.

$$g(x) > 0, if \; y_i = 1$$
$$g(x) < 0, if \; y_i = -1 \quad (8)$$

For linearly non-separable cases, one can introduce slack variables ξ and accordingly, the discriminant function is defined by

$$y_i (w.x_i + b) \geq 1 - \xi \geq 0 \quad (9)$$

measure the deviation of a data point from optimal hyper plane. SVM are designed by minimizing

$$\Phi(w, \xi) = \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \xi_i$$
$$subject\ to : y_i (w.x_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \quad (10)$$

The optimization problem is solved in a dual problem:

$$W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j (x_i . x_j)$$
$$Subject\ to : (1) 0 \leq \alpha_i \leq C, i = 1,...., m \quad (11)$$
$$(2) \sum_{i=1}^{m} \alpha_i y_i = 0$$

where α_i are the Lagrange coefficients.

The linear SVMs can be readily extended to nonlinear SVMs where more sophisticated decision boundaries are needed. This is done by applying the kernel trick, simply replacing every dot product $(x_i . x)$ in linear SVMs by a nonlinear kernel function $K(x_i . x)$, which satisfies Mercer's Theorem, called the mapping function. At each step, w is calculated by training a SVM, then remove a weak feature measured by its weight value w_i. However, a weak feature may still be useful when used with other features.

3. 2 . Minimizing the Essential Set

In this section, we describe the proposed MES feature selection method which uses the essential set sizes to evaluate the importance of a set of features. A minimum essential set is the smallest set that can correctly classify all training samples through the ranking error calculating. Since the 1-NN rule is directly related to the number of training samples involved, the size of the essential set is closely related to structural risk minimization (SRM) and thus the generalizing ability. The small sized set of learning data is to find the decision function which can minimizes the risk on test data The guaranteed risk can be derived through the bounds on the actual risk [14].

For the given training set, let *E[R(f)]* be the expection of the probability of error taken over both training and test data for an optimal function *f*

constructed on training samples of size m. Let N_m denote the size of the essential formed on the basic of training samples of size m. The following inequality holds true,

$$E[R(f)] \le \frac{E(N_m)}{m} \qquad (12)$$

Apparently, the removal of a sample $x_i \notin$ MES from the training set will not change the size of essential size. Therefore, the number of errors by leave one out method does not exceed the size of the MES, that means, the largest error rate for training data using leave one out method is Nm/m.

From equation (12) we conclude that the generalization ability of the indicator function constructed on the basis of the MES depends on the size of the MES. Minimizing the size of the MES on the basis of empirical data leads to minimizing the structure risk *R (f)*. For two feature sets with the same size, we can create two minimum reference sets for zero training errors. The feature set with a smaller MES is expected to have better generalization ability, as fewer training samples are used for constructing the classifier. Thus, the proposed MES method seeks for the feature subset that needs smallest MES for classification. We first describe the procedures to find a MES. Starting with an empty set, we update a reference set by adding the closest samples between classes until all training samples are correctly classified through 2-Norm linear classification. In the worst case, all training samples are included into the essential set .For calculating distances between samples on different classes, the Euclidean distance $d(x_i, x_j)$ is used.

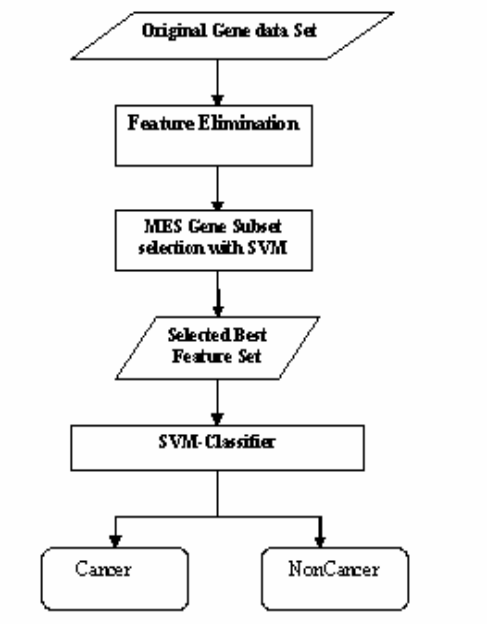

Fig.1. The procedures of MES-SVM

3.1. MES Identified Algorithm

I = set of selected samples=null
ERR (*I*):classification error of training samples in *I.*
d: ranked distances calculated from samples of between classes.
d_k: k element in d
Step1: calculate the distance d(xi,xj) for simple from two classes .i.e. , y_i=1 and y_j=-1
Step2: sort the distance from the smallest to the largest and store the ranked distance in d.Set k=1.
Step3: repeat
Find i and j which is related to $d\ (xi,\ xj) = d_k$
If $\{i, j\} \not\subset I$
Update $I \leftarrow I \cup \{i, j\}$
End if
k=k+1
until *ERR(I)* = 0
return (*I*)
The final set *I* is essential subset.

Next ,assume that the number of features to be selected is k, our method randomly chooses a set of k features and swaps one feature at a time between the selected feature set and remaining feature set. For each feature combination, EMS Identifier algorithm is executed to obtain the essential set. If the size of the MES for the selected set is smaller than that before swapping ,the swapping is accepted, otherwise, the feature set remains the same. We repeat this process for all features. The smallest number of a essential set is considered as the best feature set.

MES Feature Selection Algorithm

k = the number of selected features
n =original number of features
N (F) =the size of feature set F
S = the size of MES
F = final feature set
SF = set of selected features
RF = set of remaining features

Step1: Randomly select k features,

$$SF = \{f_1, f_2, f_3, \ldots\ldots, f_k\}$$

$$RF = \{f_{k+1}, f_{k+2}, \ldots\ldots\ldots, f_n\}$$

Step2: Search possible k features with smallest set.
Perform MES identified with feature set *SF*,
F =SF;
S =N (SF)
For i=1 to k
For j=k+1 to n
swap *fi* in *SF* and *fj* in *RF*

$S_1 = N(SF)$
If $S_1 < S$,
accept the swap $(S=S_1, F=SF)$
end if
end j
end i
return (F)

The best feature set is smallest MES that saved in *F*.

Computationally, the essential feature selection method executes Identified algorithm k*(n-k) times. Each time, one feature in selected feature set will be replaced by different feature. The new feature set is then evaluated as a whole, instead of evaluating one feature at a time in the RFE method. For better results, the search process can be repeated several times with randomly. Alternatively, one can run the algorithm just once by using a starting feature subset created by feature elimination algorithm. The major difference between MES method from other methods are(1)our method evaluates the importance of a group of features (2) MES evaluates feature set which are directly tied structural risk minimizing principle and thus good generalization while training with SVM.

4. Experimental Results

In this sub section, two criteria of evaluations effectively , number of selected genes and predictive accuracy, overall performance is substantially enhanced. By using leave-one-out validation, no need to be normalized for each gene.

4.1 . Datasets Description

We present results on some data sets which consist of a matrix of gene expression vectors obtained from DNA micro-arrays for a number of patients. The first set was obtained from cancer patients with two different types of leukemia. Although the separation of the data is easy, the problems present several features is difficulty, including small sample sizes and data differently distributed between training and test set .For colon cancer data set ,the samples can be well separated. The two subclasses studied in the lymphoma data sets are hardly separable as observed. In this work, the data sets are randomly split into independent training and test sets and applied linear SVM on them.

(1) Leukemia cancer dataset

Leukemia dataset consists of 72 samples: 25 samples of acute myeloid leukemia (AML) and 47 samples of acute lymphoblastic leukemia (ALL). The source of the gene expression measurements was taken form 63 bone marrow samples and 9 peripheral blood samples. Gene expression levels in these 72 samples were measured using high density oligonucleotide microarrays [8].Among them,38 out of 72 samples were used as training data and the remaining were used as test data in this paper. Each sample contains expression level of 7129 genes.

(2) Colon cancer dataset

Colon dataset consists of 62 samples of colon epithelial cells taken from colon-cancer patients. Each sample contains 2000 gene expression levels. 40 of 62 samples are colon cancer samples and the remaining are normal samples. Each sample was taken from tumors and normal healthy parts of the colons of the same patients and measured using high density oligonucleotide array out of 62 samples were used as training data and the remaining were used as test data .

(3) Lymphoma cancer dataset

Lymphoma data sets cell diffuse large cell lymphoma (B-DLCL) is a heterogeneous group of tumors,4026 genes containing based on significant variations in morphology and clinical presentation[11].Gene expression profiling has revealed two distinct tumor subtypes of B-DLCL: germinal center B cell-like DLCL and activated B cell-like DLCL. This data sets contains 77 tissue samples,58 are diffuse large B-cell lymphomas(DLBLC) and remaining 19 samples are follicular lymphomas(FL).

4.2. Results

Since MES and RFE are selected features through SVM classifier respectively, we compare classification performance and selected feature. In some existing methods, all features have been identified as cancer related genes by such algorithms. In this method, for case of AML/ALL, zero error rate is accepted when 64 genes are selected, actually 45 genes are identified as essential features. In the case of colon cancer, we can discover 13 genes effectively from 2000genes.For Lymphoma data sets, just 22 genes are needed for cancer identification. In this method, if the selected features is less than 30, the MES-SVM performs the better accuracy. But if the selected features larger than 30,both methods are comparable. Thus, the accuracy is raised in colon cancer and Lymphoma cancer case. Figure 2 shows the number of selected features used by different feature selection method. With less than 30 features to use, MES and RFE methods are comparable. With more than 30 features, MES performs the better accuracy than other feature selection. The probability of generalization error, using selected feature is small, MES yields higher accuracy than the Feature Elimination with SVM. We randomly selected a training data set and a test set on all data sets. In practically, small number of selected features is preferred to overcome overfitting problems. The MES method with SVM gives the better classification results. Starting from the initial

population, it can reduce many of those irrelevant genes and finally terminates with a very small number of relevant genes. If SVM classification without any feature selection method, the accuracy rate will be lower because of redundant features.

Table.1. Best results obtained by MES-SVM

Data set	Best training accuracy	Sample per class	Minimum number of selected genes
Leukemia	98.02%	ALL&AML	45
Colon	99.09%	tumor or normal	13
Lymphoma	98.85%	relapse & non relapse	22

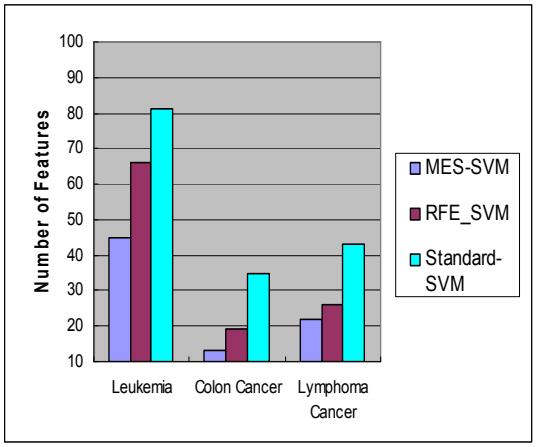

Figure. 2. Comparison of number of selected features

5. Conclusion

We have presented a method to perform feature selection for SVMs. This method is computationally feasible for high dimensional data sets and compared to existing feature selection methods. SVMs lend themselves particularly well to the analysis of broad patterns of gene expression from microarray data. It can easily deal with a large number of features (thousands of genes) and a small number of training samples. This work has integrated with feature selection method in a single consistent framework. By improving stability and accuracy, this algorithm can predict and classify the cancer types more reliably and also has the potential to identify more cancer-related genes. In the future, we plan to further explore the relationship between our method and SVM to devise the best classification result. When nonlinear kernels are used, the algorithm performs feature selection in a high dimensional space of the dual variables that depends on a small number of kernel functions In addition, MES-SVM can be applied to classification problems in very large dimensional input spaces and hopefully is a valuable addition to the methods of machine learning.

Acknowledgements

This research was supported by SDRC (Software Development and Research Center) in University of Computer Studies, Yangon. SDRC is a Research Center designated by Myanmar Science and Engineering Foundation and Ministry of Science & Technology.

References

[1] (Alizadeh, 2000) Distinct types of diffuse large B- cell lymphoma identified by gene expression profiling. Ash A. Alizadeh et al, Nature, Vol. 403, Issue 3,February, 2000.

[2] (Bannette,2003) J. Bi, K. P. Bennett, M. Embrechts, C. M. Breneman, and M. Song. Dimensionality reduction via sparse support vector machines. Journal of Machine Learning Research, March 2003.

[3] Blake,C.,Keogh,E., Merz,CJ:UCI repository of machine learning databases. Technical report Deparment of Information and Computer Science,University of California, Irvine,CA (1998).

[4] Bradley, P. S. and Mangasarian, O. L. (1998). Feature selection via concave minimization and support vector machines. Proc. 13th ICML, 82-90, San Francisco, CA.

[5] (Brown, 2000) Knowledge-based analysis of microarray gene expression data by using support vector machines, Michael P. S. Brown, William Noble Grundy, David Lin, Nello Cristianini, Charles Walsh Sugnet, Terrence S. Furey, Manuel Ares, Jr., and David Haussler. PNAS, Vol. 97, no. 1: 262–267, January, 2000.

[6] Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. JMRL special Issue on variable and Feature Selection 3, 1157-1182.

[7] H. Stoppiglia and G. Dreyfus. Ranking a random feature for variable and feature selection. *Journal of Machine Learning Research, Special Issue on Variable/Feature Selection*, 2003. In this issue.

[8] (Kohavi, 1997) Wrappers for feature subset selection. Ron Kohavi and George John. In Artificial Intelligence journal, special issue on relevance, Vol. 97,

[9] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. Sugnet, T. S. Furey, Jr. M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data using support vector machines. *Proceedings of the National Academy of Sciences of the United States of*

America, 97(1):262–267, 2000.

[10] P.Broberg, ”Statistical Methods for Ranking Differentially Expressed Genes” *Genome Biology*,vol.4,no.6,p.R41,2003.

[11] Paetz,J. “Feature selection for RBF networks” Neural Information Processing, 2002. ICONIP apos;02. Proceedings of the 9th International Conference on Volume 2, Issue , 18-22 Nov. 2002 Page(s): 986 - 990 vol.2

[12] Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *JMLR special Issue on variable and Feature Selection* 3, 1371-1382.

[13] Weston, J. Elisseeff, A. Scholkopf, B. and Tipping, M.(2003) Use of the zero-norm with linear models and kernel methods. JMLR special Issue on variable and Feature Selection 3, 1439-1461.

[14] (Weston, 2000-a) Feature Selection for SVMs. J. Weston, S. Muckerjee, O.Chapelle, M. Pontil, T. Poggio , and V. Vapnik. Submitted to NIPS 2000.

[15] V.Vapnik and O.Chapelle. Bounds on error expection for support vector machines.Neural Computation,12(9),2000

[16] Xiong, H. and Chen, X. (2006). Kernel-Based Distance Metric Learning for Microarray Data Classification. *BMC Bioinformatics*, 7:299.

[17] X.Zhang and W.H. Wong,”Recursive Sample Classification and Gene Selection Based on SVM:Method and Software Description”,Dept. of Biostatistics,Harvard School of Public Health,2001.