Gene Selection and Classification Using Non-linear Kernel Support Vector Machines Based on Gene Expression Data

Zhang Qizhong

College of electrical engineering Zhejiang university
School of automation Hangzhou dianzi university
Hangzhou, Zhejiang Province, China

zqz@hdu.edu.cn

Abstract - **In microarray-based cancer classification, feature selection and classification method is an important issue owing to large number of variables (gene expressions) and small number of experimental conditions. For disease diagnosing, classifiers' performance has direct impact on final results. In this paper, a new method of gene selection and classification by using non-linear kernel support vector machine(SVM) based on recursive performance elimination(RFE) is proposed. It is demonstrated experimentally that our method has better comprehensive performance than other linear classification methods, such as linear kernel support vector machine and fisher linear discriminant analysis (FLDA), also better than some non-linear classification methods, such as least square support vector machine(LS-SVM) using non-linear kernel. In the experiments, besides test set, leave-one-out algorithm is also used to test the classifiers' generalization performance. AML/ALL dataset and hereditary breast cancer dataset are used, which are available on internet.**

KEYWORDS: **Data classification, Support vector machine, Gene selection**

I. INTRODUCTION

One of the important breakthroughs in experimental molecular biology in recent year is microarray technology. It allows scientists to monitor the expression of genes on a genomic scale [2]. Such a technology increases the possibility of cancer classification and diagnosis at gene expression level. Many factors may affect the outcome of the analysis. One of them is the huge number of genes included in the original data. If dimension of the samples is too high, low calculating rate and over fitting appear frequently when training classifiers. Lots of methods have been taken to get over this problem, of which extracting small subset of highly discriminant genes is a remarkable one [5] [6]. In [6], feature selection and classification based on linear kernel SVM recursive feature elimination (RFE) was used. Comparing with some other gene selection and classification methods it has better performance. But linear kernel has its intrinsic limitation, when strong noisy or non-linear data samples were included in dataset, it may have a bad result. Non-linear kernel SVM can deal with this problem, but its complexity of calculation is bored. In order to settle this problem, some other eclectic steps are adopted, and satisfied results are gotten finally. Comparing with three other methods in our paper, non-linear kernel SVM method presents the best comprehensive performance whether on test data or on training data. In the end, we make an analysis on the experiment results of LS-SVM, linear kernel SVM, FLDA and non-linear kernel SVM.

The organization of this paper is as follows. A brief description of SVM is given in Section 2. Feature selection methods are outlined in Section 3. The main classification algorithms mentioned in this paper are described in Section 4. Experimental results are reported in Section 5. Analysis on the results are presented in Section 6.

II. SVM DESCRIPTION

In our method, SVM is used to perform feature selection and classification. Here, a brief introduction to original SVM algorithm was given.

SVM is a new type of learning algorithm, originally introduced by Vapnik and co-workers [15] [16], and successively extended by many other researchers. SVM's remarkable robust performance with respect to sparse and noisy data makes them first choice in a number of applications from text categorization to protein function prediction.

When being used for classification, SVMs separate a given set of binary labelled training data with a hyper-plane that is maximally distant from them (known as 'the maximal margin hyper-plane'). In cases no linear separation is possible, SVMs can work in combination with the technique of 'kernels', that automatically realizes a non-linear mapping to a feature space. The hyper-plane found by the SVM in feature space corresponds to a non-linear decision boundary in the input space [5] [12].

Suppose that we are given a training data set of n data points $\{x^i, y^i\}_{i=1}^m$, where $x^i \in R^d$ is the i-th input vector and $y^i \in R$ is the corresponding i-th target. For binary classification problem, y^i takes only two possible values $\{-1,+1\}$. In kernel designs, we employ the idea to transform input patterns into reproducing kernel Hilbert space (RKHS) by a set of mapping functions $\phi(x)$. Let us denote the reproducing kernel in RKHS as $K(x,x')$, which is defined as

$$K(x,x') = \phi(x) \cdot \phi(x') \quad (1)$$

In the RKHS, a classification is performed. The discriminant function takes the form

$$y(x) = \sum_{i=1}^{n} \omega \cdot \phi(x) + b \tag{2}$$

where ω is the weight vector in RKHS, and $b \in R$ is bias term. Equation (2) represent a hyper-plane in RKHS, if we get ω and b [15] [16].

In all these hyper-planes, there is one that can maximize the distance from nearest data point to themselves, which is the optimal hyper-plane. To find the optimal hyper-plane, we transform the problem to the dual problem as follow

$$y^i[\langle \omega, \phi(x^i) \rangle + b] \geq 1 - \xi_i, i = 1,...m \tag{3}$$

where $\xi_i \geq 0$,the generalized optimal separating hyper-plane is determined by the vector ω , that minimizes the function,

$$\Phi(\omega, \xi) = \frac{1}{2}\|\omega\|^2 + C\sum_i \xi_i \tag{4}$$

subject to constraint (3) (where C is a given value). The solution to the optimization problem of Equation (4) under the constraints (3) is given by saddle point of Lagrangian,

$$\begin{aligned} \Phi(\omega, b, a, \xi, \beta) &= \frac{1}{2}\|\varpi\|^2 + C\sum_i \xi_i \\ &- \sum_{i=1}^{m} \alpha_i (y^i[\omega^T \phi(x^i) + b] - 1 + \xi_i) - \sum_{i=1}^{m} \beta_i \xi_i \end{aligned} \tag{5}$$

where α , β are the Lagrange multipliers. The Lagrangian has to be minimized with respect to ω , b , x and maximized with respect to α , β . The dual problem is given by,

$$\begin{aligned} &Q(\alpha) = \\ &\arg\min_{\alpha} \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y^i y^j \phi(x^i)\phi(x^j) \\ &- \sum_{k=1}^{m} \alpha_k \end{aligned} \tag{6}$$

s.t.

$$0 \leq \alpha_i \leq C$$

$$\sum_{i=1}^{m} \alpha_i y^i = 0 \tag{7}$$

If we get optimal α^* , then we can get ω^* from the equation below.

$$\langle \omega^*, x \rangle = \sum_{i=1}^{m} \alpha_i y^i K(x^i, x) \tag{8}$$

b^* can be gotten from constraint (3) by using support vectors (SV) [17] [18], where α^* , ω^* , and b^* are the parameters of optimal hyper-plane. And the optimal separating hyperplane in the feature space is given by

$$f(x) = \mathrm{sgn}(\sum_{i \in SVs} \alpha_i^* y^i K(x^i, x) - b^*) \tag{9}$$

III. Gene Selection Using Non-linear Kernel Support Vector Machines

Gene selection is a key precondition for the classification. In this section, we will explain why it is a so important in the whole process, and try to find a good method to deal with this problem.

The number of features captured in gene expression data is very large. Data over fitting arises when the number of features is large (in our case thousands of genes) and the number of training patterns is comparatively small (in our case a few dozen patients). In such a situation, one can easily find a decision function that separates the training data (even with a linear decision function) but will perform poorly on testing data.

We start with a dataset S consisting of m expression vector $\mathbf{x}^i = \{x_1^i, ...x_n^i\}, 1 \leq j \leq m$, where m is the number of tissue or cell samples and n is the number of genes measured. Each sample is labelled with $Y \in \{+1, -1\}$ (e.g. cancer vs. normal). Recursive Feature Elimination (RFE) has three steps: 1) Train the classifier; 2) Compute the ranking criterion for all features; 3) Remove the feature with smallest ranking criterion. Linear kernel SVM-RFE has been used on gene expression data [6], it was demonstrated experimentally that the genes selected by this technique yield better results than the feature selection algorithms in [6]. In their algorithm, they use SVM-RFE to rank all genes in Leukemia dataset (7129 genes, 72 patients), 3 hours is spent on the operation. If non-linear kernel were used in their algorithm, much more time is needed.

In order to accelerate the feature selection process, T-Test is used to reduce the number of candidate genes. We perform a two-sample T-test on each gene to see if it is differentially expressed between normal and mutant samples. In a training set, we take all samples whose Y is equal to -1 as first group, and the others as second group. If number of samples in first group is m, and number of samples in second group is n, the T-statistic follows a t-distribution with n+m-2 degrees of freedom and is given by (10) on each gene.

$$T = \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{nS_X^2 + mS_Y^2}{n+m-2}}\sqrt{\frac{1}{n}+\frac{1}{m}}} = \frac{(\bar{X} - \bar{Y}) / \hat{\sigma}_{jo\,\mathrm{int}}}{\sqrt{\frac{1}{n}+\frac{1}{m}}} \quad (10)$$

where $\hat{\sigma}_{jo\,\mathrm{int}}$ is an unbiased estimator for the standard deviation and has computational formula:

$$\hat{\sigma}_{jo\,\mathrm{int}}^2 = \frac{nS_X^2 + mS_Y^2}{n+m-2} = \frac{\sum_1^n X_i^2 - n\bar{X}^2 + \sum_1^m Y_i^2 - m\bar{Y}^2}{n+m-2} \quad (11)$$

$$S_X^2 = \sum_1^n X_i^2 - n\bar{X}^2, S_Y^2 = \sum_1^m Y_i^2 - m\bar{Y}^2$$

Then, we can select genes by giving a confidence level. Because gene expression data may have down-regulated or up-regulated, so a two tails T-Test is adopted.

After a smaller number of genes are selected, we use non-linear kernel support vector machine based on recursive feature elimination to do the left work. We describe it in Matlab code fashion.

Inputs:

Training examples

X0 = [$x_1, x_2, \ldots, x_k, \ldots, x_\ell$]T

Class labels

y = [$y_1, y_2, \ldots, y_k, \ldots y_\ell$]T

Initialize:

Subset of surviving features

s = [1, 2, … n]

Feature ranked list

r = [ ]

Gene Ranking

Repeat until s = [ ]

Restrict training examples to good feature indices

X = X0(:, s)

Train the classifier

α = non-linear kernel SVM-train(X, y)

Compute the ranking criteria $\frac{1}{2}\alpha^T H(-i)\alpha$, $H(-i)$ is a matrix with elements $y_h y_k K(\mathbf{x}_h(-i), \mathbf{x}_k(-i))$, in which $\mathbf{x}_h(-i)$ means the i-th row of $\mathbf{x}_h$ is equal to 0. If RBF kernel is used, $K(\mathbf{x}_h, \mathbf{x}_k) = \exp(-\gamma\|\mathbf{x}_h - \mathbf{x}_k\|)$, γ is selected by us

$$\mathrm{DJ(i)} = \frac{1}{2}\alpha^T H(-i)\alpha \quad (12)$$

Find the feature with smallest ranking criterion

f = argmin(DJ)

Update feature ranked list

r = [s(f), r]

Eliminate the feature with smallest ranking criterion

s = s(1:f-1, f+1:length(s))

Output:

Feature ranked list r.

In our experiments, γ is selected as 1.

IV. Classification

In the process of classification, classification algorithm is the main point. The algorithms used in this paper are list below

A. *Linear SVM & RBF SVM*

All above we explained is original SVM. Via above description, we have been able to construct two kinds of classifiers. If using Radial basis functions (RBF) as our classifier's kernel function, we call it RBF SVM, it's our non-linear kernel SVM. If using linear functions as classifier's kernel function, we call it Linear SVM.

B. *Least square SVM (LS-SVM)*

In the last years, many researchers have extended SVM in many aspects. LS- SVM is a notable one. It was proposed by J.A.K. Suykens [9] as an interesting reformulation of the standard SVMs. As for the algorithm designs, J.A.K. Suykens employed the conjugate-gradient methods to solve the KKT system of linear equations, in which the solution to two linear systems with the order of the number of training samples are required. The well-known SMO algorithm for SVM has been extended to LS SVMs by Keerthi and Shevade [10]. When use non-linear kernel, its training is more faster than original SVM [19].

Contrast to original SVM's constratint (3) and equation (4), LS-SVM's primal problem are equation [13] and [14]. And its dual problem's deducation for optimal process is mostly the same as original SVM.

$$\min_{\omega,\xi,b} P(\omega, b, \xi) = \frac{1}{2}\|\omega\|^2 + \frac{C}{2}\sum_{i=1}^m \xi_i^2 \quad (13)$$

$$y^i - (\omega \cdot \phi(x^i) + b) = \xi_i, \forall i \quad (14)$$

In our experiment, in order to compare LS-SM with our RBF kernel SVM, we use RBF function as the kernel function of LS-SVM.

C. FLDA

FLDA (fisher linear discriminant analysis) is first applied by Barnard (1935) [1] at the suggestion of Fisher (1936) [4], It is based on finding linear combinations $\mathbf{X}\alpha$ of the gene expression levels $\mathbf{X} = (x^1, ..., x^m)$ with large ratios of between-group to within-group sums of squares. For an $n \times p$ learning set data matrix $\mathbf{X}$, the linear combination $\mathbf{X}\alpha$ of the columns of $\mathbf{X}$ has a ratio of between-group to within-group sums of squares given by $\alpha' B\alpha / \alpha' W\alpha$, where B and W denote the $p \times p$ matrices of between-group and within-group sums of squares and cross-products. The extreme values of $\alpha' B\alpha / \alpha' W\alpha$ are obtained from the eigenvalues and eigenvectors of $W^{-1}B$. The matrix $W^{-1}B$ has at most $s = \min(K-1, p)$ nonzero eigenvalues, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_s$, with corresponding linearly independent eigenvectors $v_1, v_2, ..., v_s$. The discriminant variables are defined as $u_l = xv_l, l = 1, ..., s$, and in particular, $\alpha = v_1$ maximizes $\alpha' B\alpha / \alpha' W\alpha$ [3].

V. Experimental Results

Based on methods mentioned in the last three sections, we begin our experiment. The whole experiment process contains 3 steps: 1) gene selection; 2) classification; 3) validation and analysis.

In gene selection process, the first question we should face up to is how many genes are needed by a classifier for a discriminant microarray data analysis. In [13] and [11], the problem is discussed. It is known that only top 5-15 genes are enough to get an excellent classification results. In our experiment, we get three groups of results based on top 5, top 10 and top 15 genes respectively.

In the process of validation and analysis, classifiers are not only tested on test data set, but also tested on the training set with leave-one-out algorithm. Leave-one-out algorithm is an important statistical estimator of the performance of a learning algorithm. Unlike empirical error, it is almost unbiased for estimating the generalization error of learning algorithm.

The datasets used in our experiment and results corresponding to them are given below

A. Breast cancer dataset:

On this dataset only leave-one-out test was done on the four classifiers.

Hereditary breast cancer dataset: It is downloaded from the web page of paper [8]. cDNA microarrays are used in conjunction with classification algorithms to show the feasibility of using differences in global gene expression profiles to separate BRCA1 and BRCA2 mutation-positive breast cancers. 22 breast tumor samples from 21 patients were examined: 7 BRCA1, 8 BRCA2, and 7 sporadic. There are 3226 genes for each tumor sample [17]. We use our methods to classify BRCA1 versus the others (BRCA2 and sporadic).

We use non-linear kernel SVM to rank genes in the dataset. The top 20 genes were list in Table 1. In top 5 genes, we found the most important two genes 1008 and 336, which was also mentioned in [17] [18].

TABLE I
STRONGEST GENES GELECTED BY NON-LINEAR KERNEL SVM-RFE IN HEREDITARY BREAST CANCER DATASET

No	Index No	Gene Description
1	498	PC4 and SFRS1 interacting protein 1
2	2761	macrophage migration inhibitory factor (glycosylation-inhibiting factor)
3	1620	ESTs
4	1008	keratin 8
5	336	transducer of ERBB2, 1
6	1446	fatty acid binding protein 5 (psoriasis-associated)
7	3010	cyclin-dependent kinase 4
8	1859	ESTs
9	556	tumor protein p53-binding protein, 2
10	1288	forkhead (Drosophila)-like 16
11	585	phytanoyl-CoA hydroxylase (Refsum disease)
12	1999	ESTs
13	809	CDC28 protein kinase 1
14	1277	nitrogen fixation cluster-like
15	523	O-linked N-acetylglucosamine (GlcNAc) transferase UDP-N-acetylglucosamine:polypeptide-N-acetylglucosaminyl transferase
16	2734	minichromosome maintenance deficient (S. cerevisiae) 7
17	3009	butyrate response factor 1 (EGF-response factor 1)
18	2699	peroxisome receptor 1
19	2893	mutS (E. coli) homolog 2 (colon cancer, nonpolyposis type 1)
20	1065	signal transducing adaptor molecule (SH3 domain and ITAM motif) 1

As shown in Table 2, in leave-one-out test, by the top 5 genes, on the training set of 22 samples, every kind of classification method get a very good result, only one or two errors was found. When we use top 10 and top 15 genes, FLDA's performance has a remarkable declination, when using top 15 genes, its error rate achieves 36.4%. But other three classification methods worked comparatively well - balanced. FLDA has been seriously affected by no-use data dimensions, it's over-fitting.

Although LS-SVM has a same error rate as RBF SVM, but it pays a lot, its number of support vectors is bigger than

RBF SVM's. In the LS-SVM's algorithm, data sparseness is a big problem, which directly lowers down its generalization performance on the test set, that will be seen in the test of AML/ALL dataset. In table 1, it seems Linear SVM has the same performance as RBF SVM, but when it goes to the true test set, faults will be found. In the experiment, Linear SVM and RBF kernel's number of SVs are almost the same, below 10. That means they all have strong performance on generalization.

TABLE II
CLASSIFIERS' PERFORMANCE ON HEREDITARY BREAST CANCER DATASET

Selection Method Classifier	Training Set(22 Samples) Leave One Out Error		
	TOP5	TOP10	TOP15
FDLA	2	1	8
Linear SVM	1	0	0
RBF SVM	1	0	0
LS-SVM	1	0	0

B. AML/ALL dataset:

On this dataset, we make comprehensive test using test set and leave-one-out method.

AML/ALL dataset is publicly available at http://www-genome.wi.mit.edu/ cgi-bin/cancer/publications/pub_paper.cgi?mode= view&paper_id=43. The microarray data contains 7129 human genes, sampled from 72 cases of cancer, of which 38 are of type B-cell ALL, 9 are of type T-cell ALL and 25 of type AML. The data are preprocessed mostly as recommended in [14]. Gene values are truncated from below at 100 and from above at 16,000; genes having the ratio of the maximum over the minimum less than 5 or the difference between the maximum and the minimum less than 500 are excluded.

Following the experimental setup in [7], the data is split into a training set consisting of 38 samples of which 27 are ALL and 11 are AML, and a test set of 34 samples, 20 ALL and 14 AML. In [6], their data analysis revealed that there are significant differences between the distribution of the training set and the test set. They tested various hypotheses and found that the differences can be traced to differences in the data sources. So it's a very good dataset to test the performance of algorithms mentioned above.

Table 3 is our top 20 genes from 7129 genes by non-linear kernel SVM-RFE. In the top 5, 4847 and 1882 exists, as in [6] [17] 18], they are the most important genes in the 7129 genes, which prove our method's good performance again. Based on the top 20 genes, we get the Table 4.

On training set, FLDA perform badly again, its errors of leave-one-out is always above 5 whether based on top 5 , top 10 or top15 genes. LS-SVM retains its good performance on the leave-one-out test again; experiment on the training set based on the top 5 genes shows it has no error. Linear SVM has a error when we use top 15 genes. RBF SVM has no error on the leave-one-out test. Obviously, RBF SVM shows the better performance this time.

When we use the four classification methods on the test set, we can see the results in the Table 4. FLDA, LS-SVM and Linear SVM all can't get a good results on the test set, whether based on top 5, top 10, or top 15, they all have one or two errors. RBF SVM show its better generalizaion performance although there are significant differences between the distribution of the training set and the test set. RBF SVM only has one error when test based on the top 5 gene, here, the Linear SVM and FLDA has one error, and LS-SVM has two errors. When test based the top 10 or top 15 genes, no error was found by using RBF SVM.

TABLE III
STRONGEST GENES SELECTED BY NON-LINEAR KERNEL SVM-RFE IN AML/ALL DATASET

No	Index No	Gene Description
1	3252	GLUTATHIONE S-TRANSFERASE, MICROSOMAL
2	6201	INTERLEUKIN-8 PRECURSOR
3	1882	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)
4	4847	Zyxin
5	1144	SPTAN1 Spectrin, alpha, non-erythrocytic 1 (alpha-fodrin)
6	6855	TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)
7	6218	ELA2 Elastatse 2, neutrophil
8	1630	Inducible protein mRNA
9	760	CYSTATIN A
10	5772	C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete alternatively spliced cds
11	4499	MYO-INOSITOL-1(OR 4)-MONOPHOSPHATASE
12	1834	CD33 CD33 antigen (differentiation antigen)
13	4107	PLECKSTRIN
14	6200	Interleukin 8 (IL8) gene
15	3320	Leukotriene C4 synthase (LTC4S) gene
16	2015	PPBP Connective tissue activation peptide III
17	2288	DF D component of complement (adipsin)
18	6405	CD36 CD36 antigen (collagen type I receptor, thrombospondin receptor)
19	1926	PTX3 Pentaxin-related gene, rapidly induced by IL-1 beta
20	3847	GB DEF = Homeodomain protein HoxA9 mRNA

TABLE IV
CLASSIFIERS' PERFORMANCE ON AML/ALL DATASET

Selection Method Classifier	Training Set(38 Samples) Leave One Out Error			Test Set Error (34 Samples)		
	TOP5	TOP10	TOP15	TOP5	TOP10	TOP15
FDLA	6	5	14	1	1	2
Linear SVM	0	0	1	1	2	2
RBF SVM	0	0	0	1	0	0
LS-SVM	0	0	0	2	2	1

VI. CONCLUSIONS

We do gene selection and classification using non-linear kernel SVM based on two benchmark gene expression dataset. We use leave-one-out algorithm and test set to evaluate our method compared with other three methods : linear kernel SVM, least square SVM and fisher linear discriminant analysis. Non-linear kernel SVM-RFE combined with T-statistic shows

very well on the feature-ranking process, its calculating rate is faster than only using linear kernel SVM as feature ranking method, and its feature ranking results is as well as or better than what linear kernel SVM produced. Then on the training set, our method shows the same performance as what LS-SVM and Linear SVM shows on breast cancer dataset, and shows better performance on AML/ALL dataset using leave-one-out algorithm. Finally, when tested on AML/ALL test set, our method shows best performance for non-linear classification, it does what the other three algorithms can't do. So non-linear kernel SVM-RFE is an effective and practical scheme, it has a stronger adaptive ability than the other three algorithms mentioned in our paper.

REFERENCES

[1] Barnard, M. (1935), The Secular Variations of Skull Characters in Four Series of Egyptian Skulls. *Annals of Eugenics*, 6, 352–371.

[2] Chen, Y., Dougherty, E. R., & Bittner, M. L. Ratio-based decision and the quantitative analysis of cDNA microarray images. Jour Biomed Optics 364--374 (1997).

[3] Dudoit, S., Fridlyand, J., and Speed, T.P. 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*. 97, 77-87..

[4] Fisher, R. A. (1936), The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7, 179–188.

[5] Furey, T.S., Cristianini, N., Duffy, N., Bedarski, D.W., Schummer, M. and Haussler, D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10), 906-914

[6] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46, 389-422.

[7] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M., Downing, J. R., Caligiuri, M. A., Bloomfeld, C. D., and Lander, E. S. (1999), Molecular Classi. cation of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286, 531–537.

[8] Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Rafeld, M., Yakhini Z., Ben-Dor, A., Dougherty, E.R., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O., Olsson, H., Wilfond, B., Sauter, G., Kallioniemi, O.-P, Borg, A., and Trent, J. 2001. Gene expression profiles in hereditary breast cancer. *The New England Journal of Medicine*. 344, 539-548.

[9] J.A.K. Suykens and J. Vandewalle, (1999) Least squares support vector machine classifiers. Neural Processing Letters 9(3) 293– 300.

[10] Keerthi, S. S. and S. K. Shevade. (2003.) SMO algorithm for least squares SVM formulations. Neural Computation, 15(2), Feb.

[11] Li, W. and Yang, Y. 2002. How many genes are needed for a discriminant microarray data analysis. 137-150. in *Methods of Microarray Data Analysis*, eds. S.M. Lin and K.F. Johnson (Kluwer Academic).

[12] Mukherjee, S., Tamayo, P., Mesirov, J., Slonim, D., Verri, A. And Poggio, T. (1999) Support vector machine classification of microarray data. Technical Report CBCL Paper 182/AI Memo 1676 M.I.T.

[13] Slonim, D., Tamayo, P., Mesirov, J., Golub, T. and Lander, E. (2000) Class prediction and discovery using gene expression data. In: *Proc. of 4th Annual International Conf. on Computational Molecular Biology (RECOMB)*, pp. 263-272, Tokyo: Universal Academy Press

[14] Tabus, I., Rissenan, J., and Astola, J., (2003) Classification and feature gene selection using the nor-malized maximum likelihood for discrete regression. *Signal Processing*, 83 pp. 713-727.

[15] V.N. Vapnik, (1998), Statistical Learning Theory, John Wiley & Sons

[16] V.N. Vapnik. (2000), *The Nature of Statistical Learning Theory* (2nd ed.), New York: Springer.

[17] Zhou, X., Wang, X., and Dougherty, E.R. (2003).Gene Selection Using Logistic Regressions Based on AIC, BIC and MDL Criteria. Biostatistics.

[18] Zhou, X., Wang, X., and Dougherty, E.R. (2003). Nonlinear-probit gene classification using mutual-information and wavelet-based feature selection. Biological Systems.

[19] W. Chu, S. S. Keerthi and C. J. Ong (2002), A note on least squares support vector machines, *Technical Report, CD-02-09,* Control Division, Department of Mechanical Engineering, National University of Singapore.