A Cascading Support Vector Machines System for Gene Expression Data Classification

Dimitris K. Iakovidis, Ilias N. Flaounas, *Student Member, IEEE*,
Stavros A. Karkanis, *Member, IEEE*, and Dimitris E. Maroulis, *Member, IEEE*

***Abstract*—Microarray technology provides the ability of monitoring the gene expression levels of thousands of genes in parallel. Gene expression data classification applies for diseases' diagnosis or prediction. We propose a novel intelligent system for the classification of multiclass gene expression data. It is based on a cascading Support Vector Machines (SVM) scheme and utilizes Welch's *t*-test for the detection of differentially expressed genes. The system was applied for the discrimination of normal and lung cancer subtypes' specimens. The overall accuracy achieved was 98.5%. The results show that the proposed system can be efficiently used for microarray data analysis.**

***Index Terms*—Classification, Gene Expression Data, Gene Selection, Microarrays, SVM**

I. Introduction

A variety of techniques have been developed by molecular biologists in order to study gene expression changes associated with biological evolution mechanisms and diseases. Microarray technology first provided the advantage of monitoring the gene expression levels of thousands of genes in parallel. Microarrays consist of large numbers of individual DNA sequences printed as spots in a systematic order on a microscope's glass. Each spot produced by a DNA microarray hybridization experiment represents the expression levels' ratio of a particular gene under two different experimental conditions [1].

Microarray technology motivated computer scientists to focus on solving biological problems such as the identification of the functional roles of the genes, the way they are organized, the way they interact and the way their expression levels are changed by various diseases. The major related research areas include the detection of differential expression, pattern discovery, class prediction and inference of regulatory pathways and networks [2].

Class prediction methods involve supervised machine learning techniques for diseases' diagnosis or prediction. This is a challenging task mainly due to the following reasons:

1. Microarray data consist of a large number of features (gene expression measurements), while the number of samples involved is disproportionaly small.
2. A significant percentage of genes is usually not associated with the problem under investigation.
3. The biochemical procedure used to produce microarrays, adds a lot of noise to the measurements.

The first two issues could lead to peaking phenomena associated with the "curse of dimensionality" [3], while the third introduces a large amount of uncertainty in our measurements, making the classification task harder. In order to remove irrelevant genes, identify the differentially expressed genes and reduce the feature space dimensions, gene selection algorithms are usually applied prior to the classification stage [2].

Several classification approaches have been proposed in the literature on microarray data including linear discriminant analysis, k-nearest neighbors (k-NN), parzen windows, decision trees, Neural Networks (NN) and Support Vector Machines (SVM) [4]-[8]. Comparative studies suggest that SVMs outperform other methods [5][9]. SVMs are remarkably robust machine learning algorithms that are based on statistical learning theory [10]. Their performance is not easily affected by sparse or noisy data, they resist overfitting and to the "curse of dimensionality".

The afore mentioned approaches have been applied to binary classification problems, such as the discrimination among normal and cancerous samples of colon, breast and ovarian cancer cases as well as the discrimination among two leukemia subtypes. The classification task becomes more complex as the number of classes increases. Multiclass classification approaches that have been proposed for microarray data classification include Multicategory SVMs for the classification of leukemia subtypes [11]; binary classifiers in conjunction with three combination scenarios, namely one-vs-one, one-vs-all and hierarchical partitioning for the discrimination of 14 common tumor types [12].

Under this framework we developed a novel system of cascading SVMs, for multiclass classification of gene expression data, which utilizes Welch's *t*-test for the detection of differentially expressed genes. The system was applied for the classification of normal and lung cancer subtypes samples [13].

The rest of this paper is organized in 3 sections. In section 2 the proposed system is described. In section 3 the

This work was realized under the framework of the Operational Program for Education and Vocational Training Project "Pythagoras" cofunded by European Union and the Ministry of National Education and Religious Affairs of Greece.

D. K. Iakovidis, I. N. Flaounas, and D. E. Maroulis are with Realtime Systems and Image Analysis Group, Department of Informatics and Telecommunications, National and Kapodestrian University of Athens, 15784 Panepistimiopolis, Ilisia, Athens, Greece (e-mail: rtsimage@di.uoa.gr).

S. A. Karkanis is with the Department of Informatics and Computer Technology, Technological Educational Institute of Lamia, Lamia 35100, Greece (e-mail:sk@teilam.gr).

0-7803-8278-1/04/$20.00 ©2004 IEEE

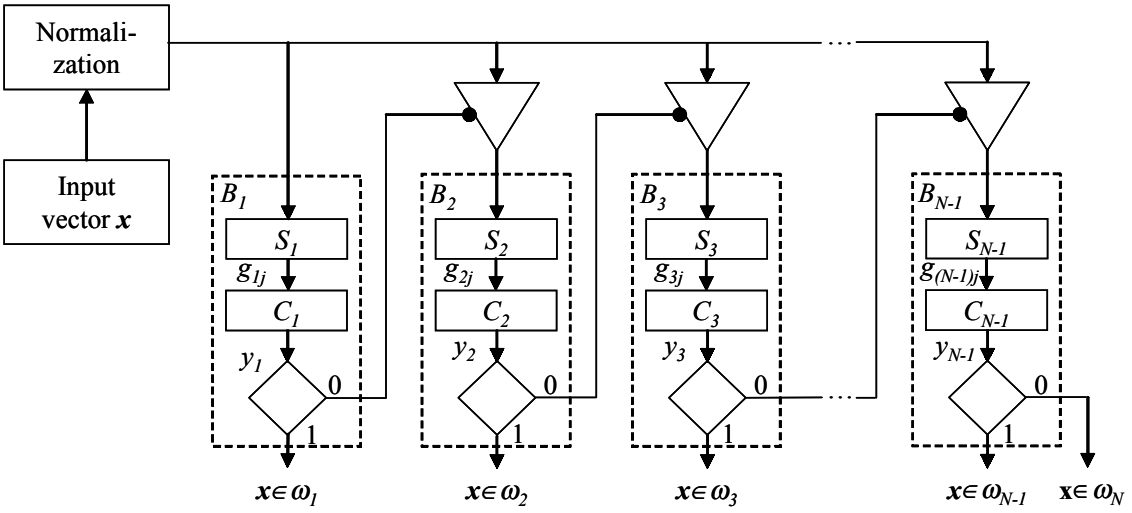

Fig. 1. Cascading SVMs system for microarray data classification.

results of the system's experimental evaluation on lung cancer data are apposed. The last section summarizes the conclusions of this study.

II. System Description

The proposed system aims to the classification of a gene expression vector $\boldsymbol{x}$ to its appropriate class ω_i, $i=1,2,...N$. The gene expression levels are normalized to conform to zero mean and unitary variance in order to obtain directly comparable sample measurements. The system implements a cascading scheme of SVM classifiers as illustrated in Fig.1. It consists of *N-1* blocks. Each block B_i consists of two modules. The first module noted as S_i, realizes gene selection and the second noted as C_i, implements classification. System's free parameters are tuned during training phase. Each block B_i is trained separately with a samples' subset X_i of the available training set X, where

$$X_i = \{x \in (\omega_i \cup \omega_h)\},\ \omega_h = \bigcup_{k>i} \omega_k \tag{1}$$

Module S_i selects a subset of v genes g_{ij}, $j=1,2...v$ which best discriminates class ω_i from class ω_h, via Welch's t-test. The number of selected genes is determined by maximizing the performance of the classification module C_i.

Presenting a vector $\boldsymbol{x}$ of unknown class to the system, module C_i is fed with the selected subset of genes, g_{ij} and outputs y_i=1 if $x \in \omega_i$ or y_i=0 if $\boldsymbol{x} \notin \omega_i$. If y_i=0, the sample enters to the next block B_{i+1}. If y_i=1, the classification task terminates and $\boldsymbol{x}$ is assigned to class ω_i. The last block B_{N-1} decides whether $\boldsymbol{x} \in \omega_{N-1}$ or $\boldsymbol{x} \in \omega_N$.

A. Welch's t-test

Welch's t-test is a statistical test that assumes unequal variances among classes and it can be applied in problems involving a small number of samples [14]. The genes are ranked based on how well they lead to a large between-class distance and a small within-class variance in the feature's space. Genes' ranking is achieved by calculating the absolute value of the t-statistic $Z(j)$ for each gene j:

$$Z(j) = \frac{m_j^i - m_j^h}{\sqrt{\frac{\sigma_j^{i\,2}}{N_i} + \frac{\sigma_j^{h\,2}}{N_h}}} \tag{2}$$

where $(m^i_j,\ \sigma^i_j)$ and $(m^h_j,\ \sigma^h_j)$ correspond to the mean and standard deviation of gene's j expression levels of the training samples that belong to ω_i and ω_h classes respectively. The number of samples belonging to each of the above classes is denoted by N_i and N_h. The larger the absolute value of $Z(j)$ the higher the expression of gene j.

B. Support Vector Machines

Let Φ be a non-linear mapping from the input space $I \subseteq \Re^n$ to the feature space $F \subseteq \Re^m$. The SVM algorithm is capable of finding a hyperplane defined by the equation

$$w\Phi(x) + b = 0 \tag{3}$$

so that the *margin of separation* is maximized. It is easy to prove [10][15] that for the *maximal margin* hyperplane,

$$w = \sum_{i=1}^{N} \lambda_i y_i \Phi^{\mathrm{T}}(x_i) \tag{4}$$

where the variables λ_i are Lagrange multipliers that can be estimated by maximizing the quantity

$$L_D = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j K(x_i, x_j) \tag{5}$$

with respect to λ_i, where the following constraints should be satisfied: $\sum_{i=1}^{N} \lambda_i y_i = 0$ and $0 \le \lambda_i \le c$, for i = 1, 2, …, N, and a given cost value c. Increasing c corresponds to a higher penalty for errors.

$K(x_i, x_j)$ is called kernel function and it is defined as the inner product

$$K(x_i, x_j) = \Phi^{\mathrm{T}}(x_i)\Phi(x_j) \tag{6}$$

Linear, polynomial, Radial Basis (RBF) and sigmoid are the most common functions used as SVM kernels. We used the RBF kernel:

$$K(x_i, x_j) = e^{-\|x_i - x_j\|^2 / \gamma} \tag{7}$$

were γ is a strictly positive constant. This kernel, usually has better boundary response as it allows for extrapolation,

and most high-dimensional data sets can be approximated by Gaussian-like distributions similar to that used by RBF networks [15].

III. Results

The experimentation presented in this study aims to the evaluation of the proposed system's performance. The dataset used has been first studied by Bhattacharjee *et al.* [13], who applied hierarchical unsupervised classification to reveal unknown adenocarcinoma subclasses. It consists of 203 samples spanning 6 different classes which correspond to normal lung specimens, Small-Cell Lung Carcinomas (SCLC), Adenocarcinomas (AC), Large-Cell Lung Carcinomas (LCLC), Squamous Carcinomas (SC) and ACs which are suspected to be extrapulmonary metastases (MAC). The number of samples per class is 17, 6, 127, 21, 20 and 12 respectively. Each sample is represented by a 12600 dimensional vector formed by the expression levels of the measured genes.

A 5-block cascading SVMs architecture was used for the 6-class classification problem. The block sequence used for the discrimination of the corresponding classes is presented in Table I.

TABLE I
System's Block Sequence for Lung Cancer Data Classification

Block	ω_i	ω_h
B_1	Normal	{SCLC, LCLC, SC, MAC, AC}
B_2	SCLC	{LCLC, SC, MAC, AC}
B_3	LCLC	{SC, MAC, AC}
B_4	SC	{MAC, AC}
B_5	MAC	AC

In each block all genes were ranked in descending significance using Welch's *t*-test. System's parameters were selected by grid search. The search parameters were the number of genes and SVM's cost *c*. Among the available genes only the 50 top-ranked were considered. Preliminary tests showed that a further increase of this number did not result in any significant increase of the classification performance. The classification performance was evaluated by adopting a Leave-One-Out (LOO) cross validation approach. LOO is commonly used when the available dataset is small providing an almost unbiased estimate of the generalization ability of a classifier [16].

Under this experimental framework the minimum number of differentially expressed genes which maximizes the classification performance of each block was determined. The classification accuracy vs. the number of genes used in blocks B_1, B_2, B_4 and B_5 is illustrated in Fig. 2, 3, 4 and 5 respectively. The diagram corresponding to the third block's performance was omitted because it reached 100% accuracy by using only the first ranked gene. Maximum accuracies are designated with vertical dashed lines within figures. The classification performances achieved as well as the number of selected genes per block are summarized in Table II. The overall accuracy of the proposed system reaches 98.5% (3 out of 203 samples were misclassified). It manages to accurately discriminate among normal specimens and different lung cancer types utilizing a rather small number of genes ranging from 1 to 40.

The results achieved are comparable with the results reported in [8]. In that study two gene selection methods namely Recursive Feature Elimination (RFE) and Univariate Association Filtering (UAF) were combined with linear and polynomial SVM, NN and k-NN classifiers for the discrimination of (i) normal - cancerous, (ii) SC - {MAC, AC} and (iii) MAC - AC specimens from the same dataset.

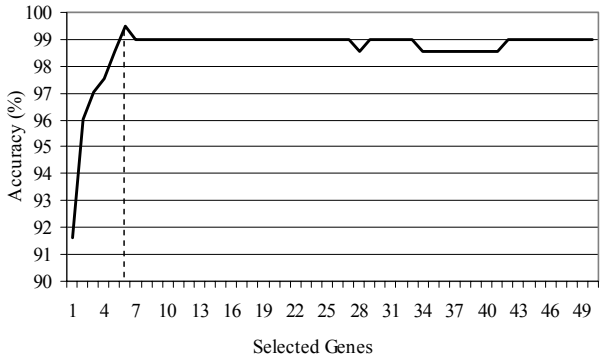

Fig. 2. Normal samples classification vs. number of genes (B_1).

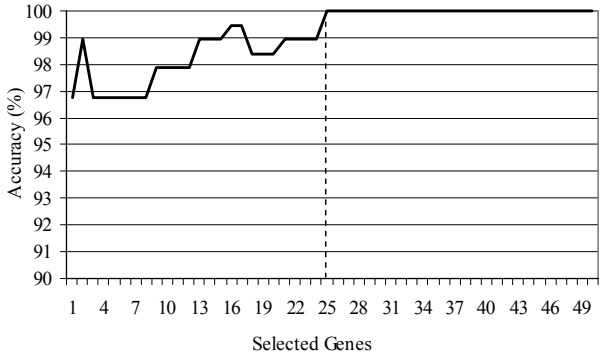

Fig. 3. SCLC samples classification vs. number of genes (B_2).

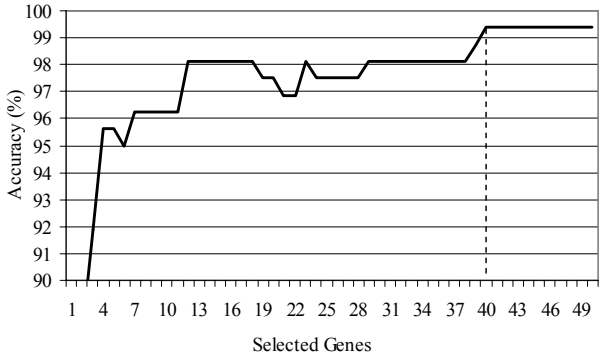

Fig. 4. SC samples classification vs. number of genes (B_4).

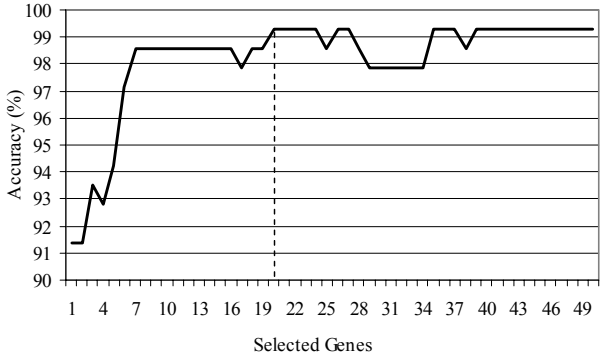

Fig. 5. MAC samples classification vs. number of genes (B_5).

These pairs of classes correspond to ω_i - ω_h pairs handled by the B_1, B_4 and B_5 blocks of the cascading SVMs system. The best results reported in [8] as well as the results of our approach are compared in Table III. In cases (i) and (ii) we achieved a comparable accuracy by using a significantly

smaller number of genes. In case (iii) the accuracy we achieved was higher by using only 14 genes more.

TABLE II
MAXIMUM CLASSIFICATION ACCURACY AND PARAMETERS USED PER BLOCK

Block	Selected Genes	Accuracy (%)
B_1	6	99.5
B_2	25	100
B_3	1	100
B_4	40	99.4
B_5	20	99.3

TABLE III
COMPARATIVE RESULTS

	Cascading SVMs System		Results reported in [8]	
ω_i - ω_h	Accuracy (%)	Selected Genes	Accuracy (%)	Selected Genes
Normal-Cancerous	99.5	6	99.8	100
SC-{MAC, AC}	99.4	40	99.6	500
MAC-AC	99.3	20	97.6	6

IV. CONCLUSIONS

In this paper we presented a novel system for the classification of multiclass gene expression data. It implements a cascading scheme of SVMs combined with gene selection modules. The proposed system was applied for the classification of lung cancer data. A 5-block cascading architecture was used for the discrimination of the six classes comprising the dataset. The results showed that the lung cancer classes could be characterized by a very small number of genes compared to the total 12600 genes involved in the experiment. The overall system's accuracy for this dataset was estimated 98.5%.

This study shows that the proposed system can be successfully used for the classification of gene expression data. A straightforward application of this system is disease diagnosis or even prediction under a medical decision support framework.

REFERENCES

[1] M. K. Deyholos, and D. W. Galbraith, "High-Density Microarrays for Gene Expression Analysis," *Cytometry*, vol. 43, no. 3, pp. 229-238, 2001.

[2] D. K. Slonim, "From patterns to pathways: gene expression data analysis comes of age," *Nature Genetics*, vol. 32, no. 12, pp. 502-508, 2002.

[3] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical Pattern Recognition: A Review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 4-37, Jan. 2000.

[4] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77-87, 2002.

[5] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey *et al.*, "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proceedings of the National Academy of Sciences USA,* vol. 97, no. 1, pp. 262-267, 2000.

[6] J. Ryu, and S. Cho, "Gene expression classification using optimal feature/classifier ensemble with negative correlation," in *Proc. International Joint Conference on Neural Networks (IJCNN'02)*, 2000, pp. 198-203.

[7] Y. Lu, and J. Han, "Cancer classification using gene expression data," *Information Systems*, vol. 28, no. 4, pp.243-268, 2003.

[8] C. F. Aliferis, I. Tsamardinos, P. P. Massion, A. Statnikov, N. Pananapazir, and D. Hardin, "Machine learning models for classification of lung cancer and selection of genomic markers using array gene expression data," in *Proc. 16th International FLAIRS Conference*, 2003, pp. 67-71.

[9] T. S. Furey, N. Christianini, N. Duffy, D. W. Bednarski, M. Schummer and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906-914, 2000.

[10] V. Vapnik, "The Nature of Statistical Learning Theory," *Springer-Verlag,* 1995.

[11] Y. Lee, and C. K. Lee, "Classification of multiple cancer types by multicategory support vector machines using gene expression data," *Bioinformatics*, vol. 19, no. 13, pp. 1132-1139, 2003.

[12] C. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R. M. Rifkin, M. Angelo *et al.,* "Molecular classification of multiple tumor types," *Bioinformatics,* vol. 17 Suppl., pp. S316-S322, 2001.

[13] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa *et al.*, "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proceedings of the National Academy of Sciences USA,* vol. 98, no. 24, pp.13790-13795, 2001.

[14] W. Pan, "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments," *Bioinformatics*, vol. 18, no. 4, pp. 546-554, 2002.

[15] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Kluwer Academic Publishers*, Boston, 1998.

[16] G. C. Cawley, and N. L. C. Talbot, "Efficient leave-one-out cross validation of kernel Fisher discriminant classifiers," *Pattern Recognition,* vol. 36, no. 11, pp. 2585-2592, 2003.