Intelligent Analysis of Genomic Measurements

I.N. Flaounas[1], D.K. Iakovidis[1], D.E. Maroulis[1], S.A. Karkanis[2]

[1] *Dept. of Informatics and Telecommunications, National and Kapodestrian Univ. of Athens, 15784 Athens, Greece, e-mail: rtsimage@di.uoa.gr*
[2] *Dept. of Informatics and Computer Technology, Technological Educational Institute of Lamia, 35100 Lamia, Greece, e-mail: sk@teilam.gr*

Abstract- In this paper we propose a methodology for intelligent analysis of genomic measurements. It is based on a sequential scheme of Support Vector Machines and it can be used for class prediction of multiclass genomic samples. The proposed methodology was evaluated using two lung cancer datasets. The results are comparable and in many cases higher to the accuracy of relevant methodologies that have been proposed in the literature.

I. Introduction

The analysis of genomic measurements is a complicated problem that can be tackled by computer scientists and statisticians. Microarray technology first provided the ability of measuring 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 [1]. The microarray glass is scanned by a special scanner capable of producing digital images (Fig. 1), which are used for the measurement of the spots' intensities.

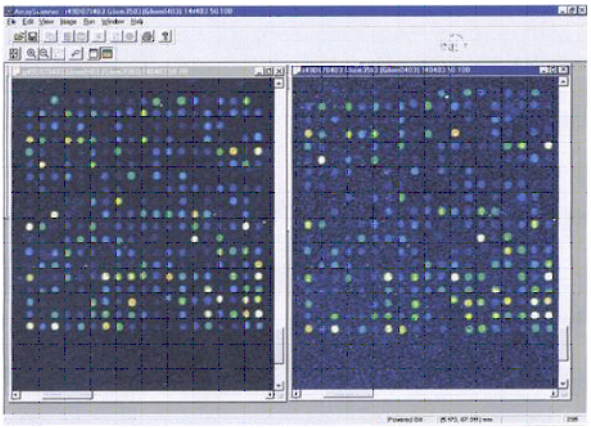


Figure 1. Microarray images.

The analysis of microarray measurements aims to the identification of the functional role 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, inference of regulatory pathways and networks and class prediction. Class prediction methods involve supervised machine learning techniques for diseases' diagnosis or prediction. This is a challenging task mainly due to three reasons: a) microarray data consist of a large number of gene expression measurements, while the number of samples is disproportionaly small, b) a significant percentage of genes is usually not associated with the problem under investigation and c) the biochemical procedure used to produce microarrays, adds a lot of noise to the measurements. Methods that have been applied for microarray measurements include linear discriminant analysis, k-nearest neighbors (k-NN), parzen windows, decision trees, Neural Networks (NN) and Support Vector Machines (SVM) [2-7]. Comparative studies suggest that SVMs outperform other methods [3][7]. SVMs are remarkably robust machine learning algorithms that are based on statistical learning theory [8]. Their performance is not easily affected by sparse or noisy data, they resist to overfitting and to the "curse of dimensionality".

In order to remove the genes that are not associated with the classification problem, identify the differentially expressed genes and reduce the dimensions of the vector space generated by the genomic measurements, gene selection algorithms are usually applied prior to the classification stage. [4] [9-13].

The performance of the proposed methodology was evaluated by following two approaches [14]. Both of them calculate the Leave-One-Out Cross Validation (LOOCV) error. The first has been widely used in the past [11-13] but Ambroise et al showed that it is biased and leads to overestimates [14]. In the second, the LOOCV procedure avoids bias.

We propose a methodology for multiclass microarray data analysis using an intelligent class prediction scheme implemented by Support Vector Machines (SVM). It utilizes a statistical ranking method for the selection of the differentially expressed genes. The methodology was applied for the classification of samples corresponding to normal and lung cancer subtypes.

II. Methodology

The proposed methodology 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 zero mean and unitary variance in order to obtain comparable sample measurements. The scheme of SVM classifiers implemented by our methodology is illustrated in Figure 2. It consists of $N-1$ blocks B_i each of which contains two modules. The first module noted as S_i, realizes gene selection and the second noted as C_i, implements classification. 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 1 if $\boldsymbol{x} \in \omega_i$ and -1 if $\boldsymbol{x} \notin \omega_i$. If $\boldsymbol{x} \notin \omega_i$, the sample enters to the next block B_{i+1} else 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 $\boldsymbol{x} \in \omega_N$.

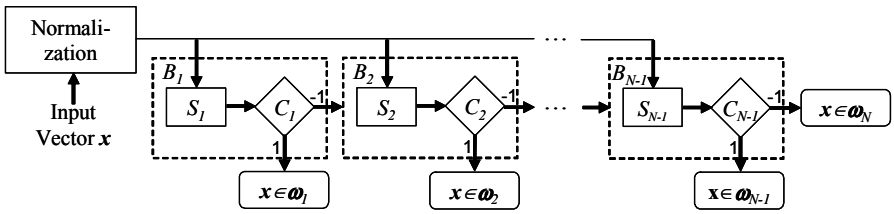


Figure 2. Multiclass microarray measurements analysis scheme.

A. Gene Selection

The statistical approach followed for the selection of the differentially expressed genes is Welch's t-test ranking. 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 [10]. 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. Classification of Genomic Measurements

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 [8][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 $i = 1, 2, \ldots, N$, 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 (of second and third order) and Radial Basis Function (RBF) are the most common functions used as SVM kernels:

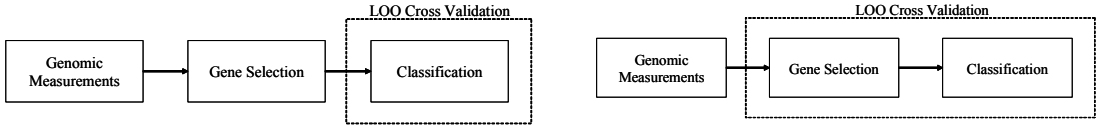
Table 1. SVM Kernels

Linear	$K(x_i, x_j) = x_i \cdot x_j$
Polynomial	$K(x_i, x_j) = (x_i \cdot x_j + 1)^p$
Radial Basis	$K(x_i, x_j) = e^{-\|x_i - x_j\|^2 / \gamma}$

were p is the order of the polynomial kernel and γ is a strictly positive constant. The linear kernel is less complex than polynomial and RBF kernels. The RBF 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].

C. System Evaluation

The two approaches followed for the evaluation of the proposed microarray measurements analysis scheme are illustrated in Fig. 3. The first calculates the Leave-One-Out cross validation error without allowance for the selection bias (LOO-1) by excluding gene selection from the LOOCV procedure. In the second the LOO procedure is external to the selection process and avoids selection bias (LOO-2).



Biased (LOO-1) Unbiased (LOO-2)

Figure 3. Leave-One-Out Cross Validation schemes.

III. Results

Two lung cancer datasets were used for the evaluation of the proposed methodology. These datasets have been firstly used for the discovery of unknown adenocarcinoma subclasses by applying hierarchical clustering [16-17]. The first dataset (D1) is comprised of 203 samples spanning 6 different classes, namely 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). Each sample consists of 12600 gene measurements. The second dataset (D2) is comprised of 65 samples spanning the first 5 of the above 6

classes. Each sample consists of 24193 gene measurements. Table 2 presents the classification sequence of each class ω_i for the two lung cancer datasets and the number of samples per class.

Table 2. Classification order for D1 and D2 and number of measured samples per class.

i	ω_i	*D1*	*D2*
1	Normal	17	5
2	SCLC	6	4
3	LCLC	20	4
4	SC	21	13
5	AC	127	39
6	MAC	12	-

The results obtained by the proposed classification scheme, for the two datasets, using the two LOO cross validation approaches and the four different SVM kernels, are summarized in Table 3.

Table 3. Results using the two lung cancer datasets.

	D1		*D2*	
SVM Kernel	*LOO-1*	*LOO-2*	*LOO-1*	*LOO-2*
Linear	96.0	82.2	100.0	70.7
Polyn. 2nd order	97.0	84.7	98.4	73.8
Polyn. 3rd order	97.0	84.7	100.0	78.4
Radial	98.5	85.2	100.0	72.3

IV. Conclusions

We presented a methodology for the analysis of genomic measurements based on an SVM multiclass classification scheme combined with gene selection modules. It was applied for the classification of lung cancer data. The results show that the accuracy of the proposed methodology is comparable and in many cases higher to the accuracy of relevant methodologies that have been proposed in the literature [6]. Moreover we have confirmed that the selection bias introduced in a performance evaluation approach (LOO-1), which is commonly used in the literature, leads to overestimated results. This conclusion is seconded by Ambroise and McLachian [14].

V. Acknowledgments

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 of Greece. It was partially funded by National and Kapodistrian University of Athens, Special Account of Research Grants.

References

[1] M. K. Deyholos, D. W. Galbraith, "High-Density Microarrays for Gene Expression Analysis," *Cytometry*, vol. 43, pp. 229-238, 2001.
[2] 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.
[3] 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.
[4] 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.
[5] Y. Lu, and J. Han, "Cancer classification using gene expression data," *Information Systems*, vol. 28, no. 4, pp.243-268, 2003.
[6] 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*, St. Augustine, Florida, USA, May 2003, pp. 67-71.
[7] 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.
[8] V. Vapnik, "The Nature of Statistical Learning Theory," *Springer-Verlag,* 1995.
[9] D. K. Slonim, "From patterns to pathways: gene expression data analysis comes of age", *Nature Genetics*, vol. 32, pp. 502-508, 2002.
[10] 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.
[11] Xiong M., Li W., Zhao J., Jin L., Boerwinkle E., "Feature (Gene) Selection in Gene Expression-Based Tumor Classification", *Mol. Genet. Metab*., vol. 73, pp. 239-247, 2001.
[12] Guyon I., Weston J., Barnhill S., Vapnic V. (2002) Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, vol. 46, 389-422.
[13] H. Zang, C.Y. Yu, B. Singer, M. Xiong, "Recursive partitioning for tumor classification with gene expression microarray data", *Proceedings of the National Academy of Sciences USA,* vol.98, pp.6730-6735, 2001.
[14] C. Ambroise, G. McLachian, "Selection bias in gene extraction on the basis of microarray gene-expression data", *Proc. Natl.Acad. Sci. USA*, vol. 99, pp. 6562-6566, 2002.
[15] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Kluwer Academic Publishers*, Boston, 1998.
[16] A. Bhattacharjee, W. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, *et al.*, "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses", *Proc. Natl.Acad. Sci. USA*, vol. 98, pp. 13790-13795, 2001.
[17] M. Garber, O. Troyanskaya, K. Schluens, S. Petersen, Z. Thaesler, M. Pacyna-Gengelbach, *et al.*, "Diversity of gene expression in adenocarcinoma of the lung", *Proc. Natl.Acad. Sci. USA*, vol. 98, pp. 13784-13789, 2001.