International Journal on Artificial Intelligence Tools
© World Scientific Publishing Company

Cascading SVMs as a Tool for Medical Diagnosis using Multi-class Gene Expression Data

ILIAS N. FLAOUNAS, DIMITRIS K. IAKOVIDIS, DIMITRIS E. MAROULIS

Dept. of Informatics and Telecommunications, National and Kapodestrian Univ. of Athens,
Panepistimiopolis, Ilissia
Athens, 15784, Greece
rtsimage@di.uoa.gr

Received (1/10/2004)
Revised (31/3/2005)
Accepted (1/6/2005)

In this paper we propose a novel Support Vector Machines-based architecture for medical diagnosis using multi-class gene expression data. It consists of a pre-processing unit and $N-1$ sequentially ordered blocks capable of classifying N classes in a cascading manner. Each block embodies both a gene selection and a classification module. It offers the flexibility of constructing block-specific gene expression spaces and hypersurfaces for the discrimination of the different classes. The proposed architecture was applied for medical diagnostic tasks including prostate and lung cancer diagnosis. Its performance was evaluated by using a leave-one-out cross validation approach which avoids the bias introduced by the gene selection process. The results show that it provides high accuracy which in most cases exceeds the accuracy achieved by the popular one-vs-one and one-vs-all SVM combination schemes and Nearest-Neighbor classifiers. The cascading SVMs can be successfully applied as a medical diagnostic tool.

Keywords: Classification; Support Vector Machines; Gene Expression Data.

1. Introduction

The breakthrough of DNA microarray technology in the last decade has motivated computer scientists to focus on biological problems such as the identification of the functional roles of the genes, the way they are organized, their interactions and the way their expression is affected by various diseases[1].

Microarrays consist of hundreds to thousands of individual cDNA or oligonucleotide sequences, often called *probes*, printed as spots usually on a glass microscope slide support. The spots are organized in an orderly and consistent way, as their location is used to identify particular genes or partial gene sequences. A mobile cDNA or mRNA target of a test subject hybridizes as it base-pairs with the probes. A special laser scanner along with an image analysis software are used for the measurement of the spots' intensities and the quantification of the gene expression levels[2,3]. The gene expression measurements acquired from a microarray comprise a large feature vector. The repetition of microarray experiments using identical cDNA probes but different targets, results in multiple feature vectors that form the so-called gene expression matrix, so that each of its rows corresponds to a particular gene or a partial gene sequence and each of its columns corresponds to a particular experiment.

Gene expression data analysis in the context of supervised classification can be applied to support medical diagnosis by providing diagnostic confirmation or clarification of unusual cases[4]. Relevant approaches to the diagnosis of various diseases include the utilization of linear discriminant analysis, *k*-nearest neighbors, parzen windows, decision trees, neural networks and Support Vector Machines (SVM)[5−8]. The studies of Brown *et al.*[7] and Furey *et al.*[9] advocate that SVMs are advantageous over other classification methods, as they are remarkably robust, their performance is not easily affected by sparse or noisy data, they resist overfitting and the "curse of dimensionality". However, SVMs have been originally designed as binary classifiers and their application for multi-class data discrimination imposes either reformulation of the SVM equations for multiple classes or the combination of multiple SVM classifiers into ensembles.

State of the art multi-class SVM schemes that have been applied for medical diagnosis using gene expression data include Multicategory SVMs[10], one-vs-one[11], one-vs-all[12], Directed Acyclic Graph (DAG)[13], Weston and Watkin's[14], and Crammer and Singer's[15]. These schemes present comparable performance in gene classification[10,11,16]. They utilize a common, constant set of genes as input in each SVM node, assuming that the various diseases correspond to separable clusters in the same gene space. Moreover, the dimensionality of this common gene space should be controlled through a gene selection process to ease the classification process, as: a) the number of the available samples is disproportionally small compared to the large number of gene expression measurements examined in a microarray experiment and b) only a small percentage of genes is differentially expressed for the various cancer types or subtypes, compared to the total number of genes involved in the experiment[1]. However, both the selection of a common gene space and the classification task become more difficult as the number of classes increases and more samples are consequently needed[17]. As only a small number of samples per class is usually available in most genomic-based diagnostic problems, the efficiency of the selection/classification task could be improved if the solution of the multi-class problem is considered as a superposition of partial two-class problem solutions. Preliminary studies have shown that such an approach can actually improve the overall accuracy of a gene expression classification system[18,19] and could possibly lead to higher classification accuracy than the accuracy provided by the standard SVM combination schemes. Therefore, a novel scheme needs to be developed for further investigation of this hypothesis.

In this paper we propose the combination of SVMs in a cascading architecture, which embodies gene selection in its structure as an attempt to meet this need. To the best of our knowledge cascading SVM architectures have not been applied for medical diagnosis using multi-class gene expression data. The proposed scheme aims at the enhancement of the diagnostic accuracy provided by the standard SVM combination schemes. It allows for the most discriminatory genes to be considered as inputs at each level of the cascading sequence by applying a gene selection criterion. Moreover the proposed architecture supports different kernel functions at each level, offering the flexibility of constructing level-specific hypersurfaces for the discrimination of the different classes.

The proposed architecture is applied for both prostate and lung cancer diagnosis using multi-class gene expression data[20,21]. The results show that it provides low classification error rates which are comparable and in most cases lower than the rates obtained by the popular one-vs-one and one-vs-all SVM combination schemes especially when a small number of genes is involved.

The rest of this paper is organized in three sections. In section 2 we describe the proposed architecture. In section 3 we appose the experimental results of the proposed approach on publicly available datasets. In the last section we discuss the results and summarize the conclusions of this study.

2. Architecture

The proposed architecture handles genomic-based medical diagnosis as a multi-class classification problem. It is capable of classifying the input gene expression vectors to their appropriate classes $\omega_i, i = 1, 2, \ldots N$. Each class consists of gene expression vectors acquired from patients suffering from the same disease or from a subtype of a particular disease. It consists of a pre-processing unit and a number of cascading blocks containing both selection and classification modules (Fig. 1).

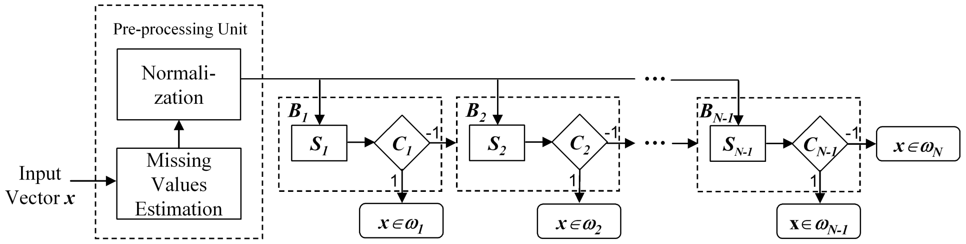


Fig. 1. Cascading architecture design.

Let N be the number of classes involved in a medical diagnosis problem. The pre-processing unit prepares the data to be inputted to the cascading blocks $B_j, j = 1, 2, \ldots N-1$ that follow. Each block contains a gene selection module S_j and a classification module C_j. The S_j module uses the output of the pre-processing unit as input. The C_j module is autonomously trained with a subset X_j of samples of the full set X of the available training samples. X_j is defined as

$$X_j = \{x \in (\omega_j \cup \Omega_h)\}, \Omega_h = \bigcup_{p=j+1}^{N} \omega_p \qquad (1)$$

The S_j module selects a subset of τ_j gene expression measurements which best discriminates class ω_i from Ω_h and maximizes the classification performance of the C_j module. Given a test vector x, the C_j module is fed with τ_j gene expression measurements, and outputs 1 if $x \in \omega_j$ or -1 if $x \notin \omega_j$. If $x \notin \omega_j$, the next block B_{j+1} will be activated to classify the test vector using the corresponding τ_{j+1} gene expression measurements. Otherwise the classification task terminates and x is assigned to the class ω_j. The last block B_{N-1} decides whether $x \in \omega_{N-1}$ or $x \in \omega_N$.

2.1. *Pre-processing unit*

The pre-processing unit is assigned to the management of missing values as well as to the normalization of the gene expression levels. Poor quality in the preparation of the cDNA targets contributes to low quality gene expression measurements as it affects the mean values and the standard deviation of the spots' intensities, their size and their contrast with the local background areas[23]. These low quality measurements are usually discarded as they lead to missing values.

A straightforward approach dealing with samples containing missing values would be to ignore these samples[24]. Unfortunately, however microarray datasets consist of a very limited number of samples and it would be a luxury to drop available data. A number of methods have been reported in the literature for coping with missing gene expression measurements. Most of these methods suggest that the missing values should be replaced by others deriving from the rest of the available data set. These include simple approaches such as the replacement of the missing values with the row-average of the gene expression matrix or more sophisticated imputation methods based on k-nearest neighbors[25], singular value decomposition[25], and Bayesian principal components analysis[26]. Acuna and Rodriguez[27] studied the effect of various missing values imputation methods, including row-average, k-nearest neighbors and median imputation, with respect to the classification accuracy and concluded that they result in comparable performance. In this paper we have adopted the row-average method as it combines both low-complexity and effectiveness[25] in the prediction of missing values.

In the sequel, the data are normalized to conform to zero mean and unitary variance using the following equation:

$$g'_{kl} = \frac{g_{kl} - \mu_l}{\sigma_l} \tag{2}$$

where g'_{kl} corresponds to the normalized gene expression level g_{kl} located in the k-th row (gene) and l-th column (sample) of the gene expression matrix, and μ_l, σ_l represent the mean and the standard deviation of the gene expression levels estimated over the l-th column. This normalization facilitates making the gene expression levels of the different DNA microarrays comparable[23].

2.2. *Gene selection modules*

The gene selection module is devoted to the selection of differentially expressed genes at each level of the cascading architecture. Guyon *et al.*[28] suggested a grouping of the gene selection methods into three main categories, namely *ranking*, *wrapper* and *embedded* methods. Ranking methods are usually used because of their simplicity, scalability, and good empirical success. They select subsets of variables independently of the chosen classifier. A standard statistical test for detecting significant changes between repeated measurements of a variable in two groups used in microarray data analysis is the t-test and several other of its variations[1]. Wrappers utilize the classifier of interest as a black box to score subsets of variables according to their predictive power. The complexity of most wrapper approaches such as Sequential Backward Selection (SBS), Sequential Forward Floating Search (SFFS) and Adaptive SFFS (ASFFS) is prohibiting for large-scale gene selection problems[29]. However, recent studies suggest that wrapper approaches based on genetic algorithms could efficiently handle large-scale gene selection[30]. Embedded methods perform variable selection in the process of training and are usually specific to given classifiers. A state of the art embedded approach that has been applied for the identification of differentially expressed genes in microarray experiments is Recursive Feature Elimination (RFE)[31].

The gene selection module of the proposed architecture could implement any of the afore mentioned gene selection methods. In order to keep the overall complexity into reasonable levels we have considered Welch's t-test[32] as an efficient gene ranking criterion. Welch's t-test is a statistical test that assumes unequal variances among classes and it can be applied to problems involving a small number of samples[33]. It is similar to the signal-to-noise ratio, one of the most widely used statistics for gene expression research which first appeared in the seminal paper of Golub *et al.*[4] but has the comparative advantage of noticing the number of the available samples per class.

The steps of the gene selection process that is followed in each selection module S_j are:

(i) Each gene g is ranked based on the t-statistic $Z(g)$:

$$Z(g) = \frac{\mu_g^j - \mu_g^h}{\sqrt{\frac{(\sigma_g^j)^2}{n_j} + \frac{(\sigma_g^h)^2}{n_h}}} \tag{3}$$

where (μ_g^j, σ_g^j) and (μ_g^h, σ_g^h) correspond to the mean and standard deviation of the expression levels of the gene g for the training samples that belong to ω_j and Ω_h classes respectively. The number of samples belonging to each of the above classes is denoted by n_j and n_h.

(ii) The genes are ordered in descending order according to the absolute value of their $Z(g)$ statistic.

(iii) The τ_j top-ranked genes are selected[4,12,24] as they lead to a large between-class distance and a small within-class variance. Alternatevely an improvement in the diagnostic accuracy could be obtained if a combination of m of τ_j genes ($m < \tau_j$) is found by permutating the τ_j genes and selecting the first m. However, this approach leads to a disproportional increase of the complexity compared to the possible marginal increase in the performance.

Trunk[34] showed that the number τ of the selected features in a classification problem cannot be arbitrarily increased when the parameters of class-conditional densities such as mean values and variances are estimated from a finite number of training samples. This is the case of the microarray datasets which comprise of a very limited number of samples. It is generally accepted that the number of inputs in a classification system should be at the most one tenth of the available training samples in order to avoid the curse of dimensionality[29,35]. Although SVMs are quite resistant to the curse of dimensionality it has been noted[36] that they are not completely insensitive to it and feature selection could indeed improve their classification performance. Furthermore, the use of a limited number of genes as features will result in the identification of reliable gene-markers for diseases.

2.3. *Classification modules*

The classification modules of the proposed architecture are based on SVMs. SVMs are machine-learning algorithms derived by Vapnik[37] in the framework of structural risk minimization, which aim at building parsimonious models, in the sense of statistical learning theory. This algorithm can be summarized as follows:

Consider an input space I of vectors $x_i, i = 1, 2, \dots v$, belonging to two classes, labeled as $y_i \in \{-1, 1\}$. 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 is capable of finding a hyperplane defined by the equation

$$w^T \Phi(x) + w_0 = 0 \tag{4}$$

so that the *margin of separation* between the two classes is maximized. It is easy to prove[37,38] that for the maximal margin hyperplane,

$$w = \sum_{i=1}^{v} \lambda_i y_i \Phi(x_i) \tag{5}$$

while w_0 is estimated by the Karush-Kuhn-Tucker (KKT) complementarity condition[38]. The λ_i variables are Lagrange multipliers that can be estimated by maximizing the Lagrangian

$$L_D = \sum_{i=1}^{v} \lambda_i - \frac{1}{2} \sum_{i=1}^{v} \sum_{j=1}^{v} \lambda_i \lambda_j y_i y_j K(x_i, x_j) \tag{6}$$

with respect to λ_i. The vectors x_i for which $0 \le \lambda_i \le c$, are called *support vectors* and c is a positive cost parameter. As the c value increases a higher penalty for errors is assigned.

The function $K(x_i, x_j)$ known as *kernel function*, is defined as the inner product

$$K(x_i, x_j) = \Phi^T(x_i)\Phi(x_j) \tag{7}$$

and should satisfy Mercer's condition[37].

Most commonly used kernel functions are the linear, the polynomial (of second and third order) and the Radial Basis Functions (RBF) as presented in Table 1. where p is the order of the polynomial kernel and γ is a strictly positive constant.

Table 1. SVM Kernels.

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

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 those used by RBF networks[38]. The separating hyperplane can be finally derived by the following equation:

$$\sum_{\forall i:0<\lambda_i<c} \lambda_i d_i K(x_i, x) + w_0 = 0 \tag{8}$$

The implementation of SVMs was based on the publicly available LibSVM library[39] which was modified to meet the needs of the proposed architecture.

2.4. *Modes of operation*

The proposed architecture supports two modes of operation: the training and the test mode. During the training mode the following parameters are automatically tuned by grid search[39]:

(i) The kernel function of each classification module.
(ii) The cost c and the γ parameters of each classification module.
(iii) The number of genes τ_j to be selected by the S_j selection module.

The parameters that maximize the overall accuracy of the system as well as the selected genes and the support vectors of each block are stored. Other SVM combination senarios such as the one-vs-one and one-vs-all use the same kernel for all of its classifiers, while the proposed cascading scheme has the advantage of using a different kernel function per classification unit. In the test mode this information is retrieved and the architecture is capable of classifying uncharacterized gene expression data for medical diagnosis.

3. Results

We performed two sets of experiments to evaluate the capabilities of the proposed architecture for medical diagnosis. The first set aims at prostate cancer diagnosis while the second aims at lung cancer diagnosis. The datasets used in the experiments are publicly available in the Stanford Microarray Database[40]. It is worth noting that these particular datasets have not been utilized for supervised genomic-based diagnosis. The results achieved with the proposed architecture are compared with the results obtained using a) standard k-Nearest Neighbor (kNN) classification approaches and b) the popular one-vs-one and one-vs-all SVMs combination scenarios, in conjunction with the modified Welch's t-test that supports multiple classes:

$$Z(g) = \sum_{j}^{N} \sum_{h \neq j}^{N} \frac{\mu_g^j - \mu_g^h}{\sqrt{\frac{(\sigma_g^j)^2}{n_j} + \frac{(\sigma_g^h)^2}{n_h}}} \tag{9}$$

where N is the total number of classes involved, (μ_g^j, σ_g^j) and (μ_g^h, σ_g^h) correspond to the mean and standard deviation of the expression levels of the gene g for the training samples that belong to ω_j and Ω_h classes respectively. The number of samples belonging to each of the these classes is denoted by n_j and n_h respectively.

The one-vs-one scheme consists of $N(N-1)/2$ SVM classifiers, each one of which is trained from samples of two classes. The outputs of these classifiers are combined using a "Max Wins" voting strategy which decides upon the class that an uncharacterized input vector x belongs to[11]. The one-vs-all scheme utilizes N classifiers, where the ith SVM is trained with all the samples of the ω_i against all the others. The one-vs-all scheme has in general comparable performance to the one-vs-one scheme but due to its higher complexity requires longer training times[11].

During the training mode for the SVM-based classification schemes (Cascading, one-vs-one and one-vs-all) the kernel functions that were tested are the linear, the 2nd and 3rd order polynomial and the RBF. The ranges of the training parame-

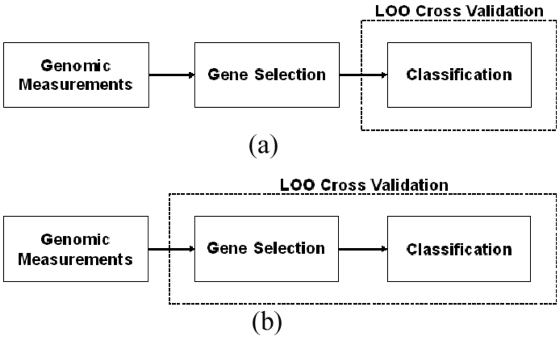


Fig. 2. Leave-One-Out Cross Validation approaches.

ters considered were 2^{-5} to 2^{15} for the cost parameter c and 2^{-15} to 2^{3} for the γ parameter as these have been proposed by Chang *et al.*[39]. The best parameters that maximize the performance for the cascading, the one-vs-one and the one-vs-all schemes were determined using grid search. The order of the blocks in the cascade model was determined based on the histopathological subclassification of carcinomas [20,21]. As regards the kNN classifiers three different k-values were tested (1,3 and 5) and the Eucledian distance was used to determine the “nearest” neighbors.

The classification performance for all of the compared classifiers was estimated in terms of classification error by using the Leave-One-Out Cross Validation (LOOCV) method [24] as the number of the available samples for both prostate and lung cancer datasets was limited. The LOOCV has been widely used for the evaluation of gene expression classification systems[41,42]. Most studies exclude gene selection from the LOOCV procedure (Fig. 2a). Recently, Ambroise *et al.*[22] showed that LOOCV should be used with the caution not to be separated from the gene selection process, as this leads to biased and overestimated results. So, in our experiments the gene selection process has been included within the LOOCV procedure in order to avoid the selection bias (Fig. 2b). Thus in every iteration of the LOOCV a new subset of genes was selected. It is worth noting that the error estimated, using both of these approaches, will be approximately the same if the number of training samples is adequate. In the case of microarray datasets, where the number of samples is limited, these two methods will result in different accuracies.

The significance of the presented classification results was assessed by permutation testing[43]. If the null hypothesis that no systematic differences in gene expression profiles exist between the classes is rejected, the results are considered significant, otherwise it can be assumed that the assignment of gene expression profiles to class labels is purely coincidental. A total of 2000 permutations was generated for each experiment to ensure that the estimate of the achieved significance level varies by less than 10% from a true achieved significance level of 0.05.

3.1. *Prostate cancer diagnosis*

Prostate cancer displays a broad range of clinical behavior from relatively indolent to aggressive metastatic disease[44]. Lapointe *et al.* published a dataset of prostate cancer gene expressions on which they applied unsupervised hierarchical clustering in order to distinguish tumors from normal samples[20]. The prostate cancer dataset is comprised of 112 samples spanning three classes, namely normal prostate tissue (41 samples), primary prostate tumors (62 samples) and lymph node metastases (9 samples). Each sample consists of 44016 gene expression measurements.

The cascading architecture built for the discrimination of these three classes consists of two blocks. The first block was assigned to decide whether an unknown input vector x is normal or cancerous. If it is classified as cancerous, the second block is activated and decides whether the cancer is primary or metastatic. A range of one to 11 genes was considered in the gene selection process based on the criterion

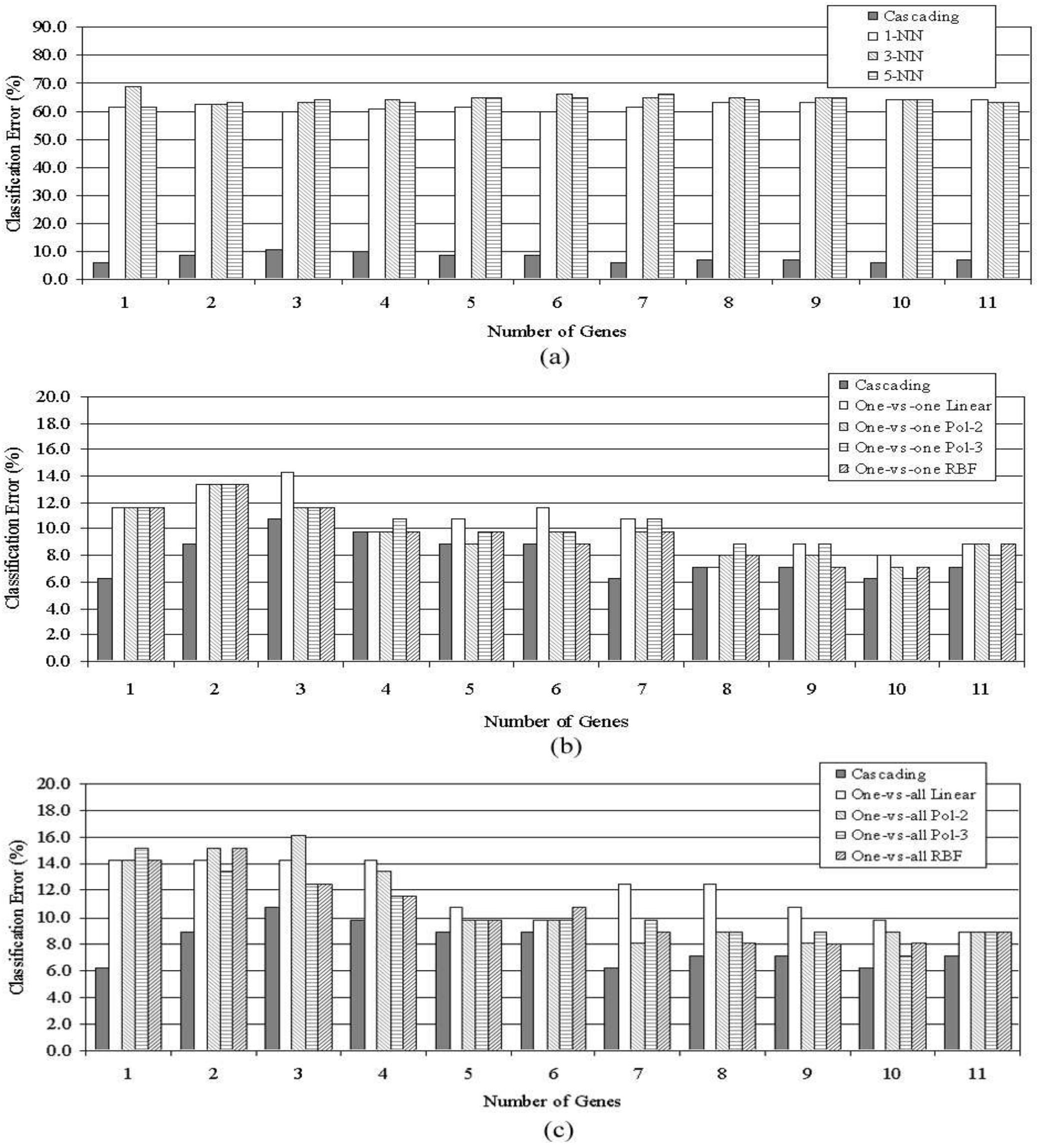


Fig. 3. Classification results for the prostate cancer dataset using a) kNN, b) one-vs-one and c) one-vs-all.

described in section 2.2. The classification performance, for the range of the selected genes, of the proposed architecture is compared with the performance of the 1-NN, 3-NN and 5-NN classifiers (Fig. 3a), the one-vs-one (Fig. 3b) and the one-vs-all SVM combination schemes (Fig. 3c). The presented results for the proposed

Table 2. Classification results (%) using the prostate cancer dataset in detail.

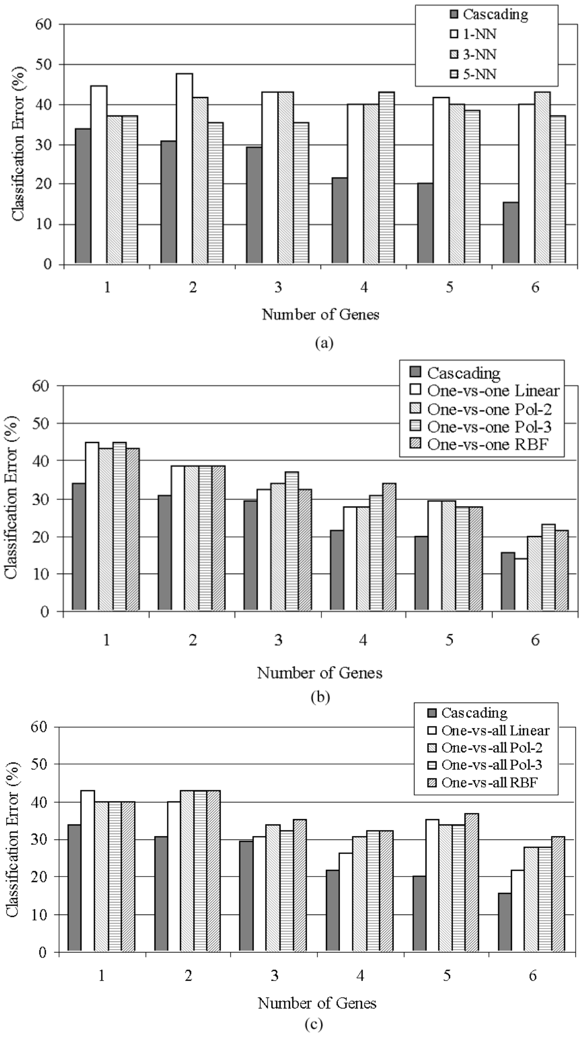
Classifier / Genes	1	2	3	4	5	6	7	8	9	10	11
Cascading	**6.3**	**8.9**	**10.7**	**9.8**	**8.9**	**8.9**	**6.3**	**7.1**	**7.1**	**6.3**	**7.1**
1-NN	61.6	62.5	59.8	60.7	61.6	59.8	61.6	63.4	63.4	64.3	64.3
3-NN	68.8	62.5	63.4	64.3	65.2	66.1	65.2	65.2	65.2	64.3	63.4
5-NN	61.6	63.4	64.3	63.4	65.2	65.2	66.1	64.3	65.2	64.3	63.4
One-vs-one Linear	11.6	13.4	14.3	**9.8**	10.7	11.6	10.7	**7.1**	8.9	8.0	8.9
One-vs-one Pol-2	11.6	13.4	11.6	**9.8**	**8.9**	9.8	9.8	8.0	8.0	7.1	8.9
One-vs-one Pol-3	11.6	13.4	11.6	10.7	9.8	9.8	10.7	8.9	8.9	**6.3**	8.0
One-vs-one RBF	11.6	13.4	11.6	**9.8**	9.8	**8.9**	9.8	8.0	**7.1**	7.1	8.9
One-vs-all Linear	14.3	14.3	14.3	14.3	10.7	9.8	12.5	12.5	10.7	9.8	8.9
One-vs-all Pol-2	14.3	15.2	16.1	13.4	9.8	9.8	8.0	8.9	8.0	8.9	8.9
One-vs-all Pol-3	15.2	13.4	12.5	11.6	9.8	9.8	9.8	8.9	8.9	7.1	8.9
One-vs-all RBF	14.3	15.2	12.5	11.6	9.8	10.7	8.9	8.0	8.0	8.0	8.9

architecture are obtained using the optimal kernel function for each classification unit as selected during training. The one-vs-one and one-vs-all approaches utilize the same predefined kernel function for each classifier, so different error rates are obtained using linear, 2nd and 3rd polynomial order and RBF kernels. Table 2 summarizes the exact classification error rates corresponding to Fig. 3.

The results for the prostate cancer dataset show that the SVM-based classifiers outperform the kNN classifiers. The permutation testing for the proposed architecture led to rejection of the null hypothesis with a p-value estimate less than 0.05. In all cases, the proposed architecture results in lower or comparable classification error rates with the one-vs-one and one-vs-all schemes. The minimum classification error obtained for the prostate cancer dataset reached 6.3% in three cases using one, seven or ten genes. The same classification error rate was obtained by one-vs-one scheme using 3rd order polynomial kernel and ten genes. So, the proposed architecture has an advantage over these schemes as it is capable of providing a better or comparable performance using fewer genes. The upper classification error bound of the cascading SVMs architecture is 10.7%, while this bound increases to 14.3% for the one-vs-one and to 16.1% for the one-vs-all scheme.

3.2. *Lung cancer dataset*

There are four main histologic subtypes of lung cancer that are regularly distinguished by tumor morphology. These include Small-Cell Lung Carcinomas (SCLC), Large-Cell Lung Carcinomas (LCLC), Squamous Carcinomas (SC) and Adeno-Carcinomas (AC)[45]. Garber *et al.*[21] published a dataset of lung cancer gene expressions of samples belonging to these subtypes. Their work was focused on unsupervised subclassification of adenocarcinoma into subgroups that correlate with the degree of tumor differentiation and patient survival. This dataset comprises of 65 samples spanning five classes as follows: 5 normal lung specimens, 4 SCLC, 4 LCLC,



(a)

(b)

(c)

Fig. 4. Classification results for the lung cancer dataset using a) kNN, b) one-vs-one and c) one-vs-all.

13 SC and 39 AC. Each sample consists of 24193 gene expression measurements.

A cascading architecture of four blocks was built for the discrimination of the five classes. The first block handles the discrimination of normal and cancerous samples, the second block separates the SCLC samples, the third block separates the LCLC samples, and the last block makes the discrimination between SC and AC samples. A range of one to six genes was considered in the gene selection process based on the criterion described in section 2.2. Comparative classification results were obtained using the proposed architecture, 1-NN, 3-NN and 5-NN classifiers (Fig. 4a), and

Table 3. Classification results (%) using the lung cancer dataset in detail.

Classifier / Genes	1	2	3	4	5	6
Cascading	**33.9**	**30.8**	**29.2**	**21.5**	**20.0**	15.4
1-NN	44.6	47.7	43.1	40.0	41.5	40.0
3-NN	36.9	41.5	43.1	40.0	40.0	43.1
5-NN	36.9	35.4	35.4	43.1	38.5	36.9
One-vs-one Linear	44.6	38.5	32.3	27.7	29.2	**13.9**
One-vs-one Pol-2	43.1	38.5	33.9	27.7	29.2	20.0
One-vs-one Pol-3	44.6	38.5	36.9	30.8	27.7	23.1
One-vs-one RBF	43.1	38.5	32.3	33.9	27.7	21.5
One-vs-all Linear	43.1	40.0	30.8	26.2	35.4	21.5
One-vs-all Pol-2	40.0	43.1	33.9	30.8	33.9	27.7
One-vs-all Pol-3	40.0	43.1	32.3	32.3	33.9	27.7
One-vs-all RBF	40.0	43.1	35.4	32.3	36.9	30.8

the one-vs-one (Fig. 4b) and one-vs-all SVM combination schemes using different kernels (Fig. 4c). The corresponding classification errors are presented in detail in Table 3.

The results show that the proposed architecture leads to lower classification error rates compared to the other classifiers using one to five genes. The permutation testing for the proposed architecture led to the rejection of the null hypothesis with a p-value estimate less than 0.05. The one-vs-one SVM combination scheme with linear kernel results in the lowest classification error for six input genes. The minimum classification error rate obtained by the cascading SVMs architecture was 15.4% whereas the maximum reached 33.9%.

4. Conclusions

We presented a novel architecture based on cascading SVMs that can be used as a reliable tool for medical diagnosis using multi-class gene expression data. It was applied for genomic-based diagnosis of prostate and lung cancer using multi-class gene expression datasets. The proposed architecture consists of a pre-processing unit and sequentially ordered blocks for the discrimination of multiple classes and it allows for the most discriminatory genes to be considered as inputs at each block by applying a gene selection criterion. This feature makes the proposed architecture particulary suitable for medical diagnosis because it allows for different gene-markers to be identified for each disease or subtype of a disease. For example, the most important gene identified for the discrimination of abnormal from normal samples in the prostate dataset is *Caveolin 1*, which is a tumor suppressor gene candidate and a negative regulator of the Ras-p42/44 MAP kinase cascade [46]. Consequently, the proposed architecture can lead to a classification model of low complexity which encompasses relevant genomic-based diagnostic information within its structure.

The order of the block sequence in a cascading classification model could be determined a) by the available knowledge provided by the medical experts on the particular diagnostic application[24], e.g. in our study we considered the histopathological subclassification of carcinomas[20,45], and b) by considering the complexity and the overall classification accuracy of the model. Studies on cascading classifiers [29,47] suggest that the classifiers should be ordered in ascending complexity. That is, the less complex classifiers should be ordered before the more accurate and complex ones. However, the proposed architecture consists only of SVM classifiers and embodies a search algorithm for the determination of their parameters, which in turn affects their complexity, e.g. a large c parameter could lead to an increase in the number of support vectors. Therefore, the ordering of the classifiers in ascending complexity is not directly feasible. Alternatively, one could experiment with all possible orders of the SVM classification blocks and decide upon the lowest classification error rates and the least overall architecture complexity.

The results of this study lead to the following conclusions regarding the SVM combination schemes:

(i) In most cases they perform better than kNN classifiers. The proposed architecture performs better than kNN classifiers in all cases.

(ii) The cascading SVM combination scheme provides low classification error rates which are comparable and in most cases lower than the rates obtained by the one-vs-one SVM combination scheme especially when a small number of genes is involved.

(iii) The cascading SVM combination scheme provides lower classification error rates in all cases compared to the one-vs-all SVM combination scheme.

(iv) The results validate the conclusions of Hsu *et al.* [11], which support that the error rates obtained using the one-vs-one and the one-vs-all SVM combination schemes are comparable to each other.

(v) The proposed architecture utilizes $N - 1$ classifiers whereas the one-vs-one SVM combination scheme utilizes $N(N - 1)/2$ classifiers and the one-vs-all SVM combination scheme utilizes N classifiers.

(vi) In the case of prostate cancer diagnosis the overall error obtained was lower than in the case of lung cancer. This could be attributed to the fact that the first dataset consists of more samples and fewer classes.

An issue that arises considering the sequential arrangement of the blocks in the cascading architecture is that an error, which may occur on a first block, will propagate to the next blocks. The impact of this “error propagation effect” to the overall classification accuracy of the cascading architecture is negligible as a sample is considered misclassified by the first time the error occurs. However, an increase of the overall processing time performance could be observed, as more blocks would be falsely enabled for the classification of the propagated samples.

The proposed cascading SVMs architecture has given promising results which, in conjunction with the decreasing cost of microarrays, advocate to its direct clinical applicability. Within our future perspectives is the incorporation of more sophisticated gene selection methods[31] and SVM based methods for tackling the problem of missing values[48]. Moreover, we consider the development of a complete standalone microarray data analysis tool for medical diagnosis which will integrate image processing and analysis methodologies [49] in addition to the cascading diagnostic architecture under a user friendly graphical environment.

Acknowledgments

This research was funded by the Operational Program for Education and Vocational Training (EPEAEK II) under the framework of the project "Pythagoras - Support of University Research Groups" cofunded by 75% from the European Social Fund and by 25% from national funds. We would also like to thank the anonymous reviewers for their helpful comments and suggestions.

References

1. D. K. Slonim, From patterns to pathways: gene expression data analysis comes of age, *Nature Genetics*, Vol. 32 (Nature Publishing Group, 2002), pp. 502–508.
2. T. Speed (ed.), *Statistical Analysis of Gene Expression Microarray Data*, (Chapman & Hall/CRC, London, 2003).
3. M. K. Deyholos, D. W. Galbraith, High-Density Microarrays for Gene Expression Analysis, *Cytometry*, Vol. 43 (Wiley InterScience, 2001), pp. 229–238.
4. T.R. Golub, et al., Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, *Science*, Vol. 286 (1999), pp. 531–537.
5. 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)* (2002), pp. 198–203.
6. 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 (2002), pp. 77–87.
7. 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*, Vol. 97 (National Academy of Sciences, USA, 2000), pp. 262–267.
8. Y. Lu and J. Han, Cancer classification using gene expression data, *Information Systems*, vol. 28 (2003), pp. 243–268.
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 (Oxford University Press, Oxford, 2000), pp. 906–914.
10. Y. Lee, and C. K. Lee, Classification of multiple cancer types by multicategory support vector machines using gene expression data, *Bioinformatics*, Vol. 19 (Oxford University Press, Oxford, 2003), pp. 1132–1139.
11. C.W. Hsu and C.J. Lin, A comparison of Methods for Multiclass Support Vector Machines, *Trans. Neural Networks*, Vol. 13 (IEEE, USA, 2002), pp. 415–425.
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 (Oxford University Press, Oxford, 2001), pp. S316–S322.

13. J.C. Platt, N. Christianini and J. Shawe-Taylor, Large margin DAGs for multiclass classification, *Advances in Neural Information Processing Systems*, Vol. 12 (MIT Press, Cambridge MA, 2000), pp. 547–553.
14. J. Weston and C. Watkins, Multi-class support vector machines, in *Proceedings of ESANN99*, (Brussels, D. Facto Press, 1999).
15. K. Crammer and Y. Singer, Ultraconservative online algorithms for multiclass problems, *Technical report, School of Computer Science and Engineering*, (Hebrew University, 2001).
16. A. Statnikov, C. F. Aliferis, I. Tsamardinos, Using Support Vector Machines for Multicategory Cancer Diagnosis Based on Gene Expression Data, (to appear)*in Proceedings of 11th World Congress in Medical Informatics* (MEDINFO '04)(2004).
17. A. M. Bagirov, B. Ferguson, S. Ivkovic, G. Saunders and J. Yearwood, New algorithms for multi-class cancer diagnosis using tumor gene expression signatures, *Bioinformatics*, Vol. 19 (Oxford University Press, Oxford, 2003), pp. 1800–1807.
18. D.K. Iakovidis, I.N. Flaounas, S.A. Karkanis and D.E. Maroulis, A Cascading Support Vector Machine System for Gene Expression Data Classification, in *2nd International IEEE Conference Intelligent Systems* (Bulgaria, Varna, 2004), pp. 344–347.
19. I.N. Flaounas, D.K. Iakovidis, D.E. Maroulis and S.A. Karkanis, Intelligent Analysis of Genomic Measurements, *13th International Symposium on Measurements for Research and Industry Applications, IMEKO* (Greece, Athens, 2004), pp. 463–467.
20. J. Lapointe, C. Li, J.P. Higgins, M. van de Rijn, E. Bair, K. Montgomery et al., Gene expression profiling identifies clinically relevant subtypes of prostate cancer, *Proceedings of the National Academy of Sciences*, Vol. 101 (National Academy of Sciences, USA, 2004), pp. 811–816.
21. M. Garber, O. Troyanskaya, K. Schluens, S. Petersen, Z. Thaesler, M. Pacyna-Gengelbach, et al., Diversity of gene expression in adenocarcinoma of the lung, *Proceedings of the National Academy of Sciences*, Vol. 98 (National Academy of Sciences, USA, 2001), pp. 13784–13789.
22. C. Ambroise and G. McLachian, Selection bias in gene extraction on the basis of microarray gene-expression data, *Proceedings of the National Academy of Sciences*, Vol. 99 (National Academy of Sciences, USA, 2002), pp. 6562–6566.
23. W. Zhang and I. Shmulevich(ed.), *Computation and Statistical Approaches to Genomics*, (Kluwer Academic Publishers, Boston, 2002).
24. S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, (Academic Press, 1999).
25. O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshiran, D. Botstein and R. B. Altman, Missing value estimation methods for DNA microarrays, *Bioinformatics*, Vol. 17 (Oxford University Press, Oxford, 2001), pp. 520–525.
26. S. Oba, M. Sato, I. Takemasa, M. Monden, K. Matsubara and S. Ishii, A Bayesian missing value estimation method for gene expression profile data, *Bionformatics*, Vol. 19 (Oxford University Press, Oxford, 2003), pp. 2088–2096.
27. E. Acuna and C. Rodriguez, The treatment of missing values and its effect in the classifier accuracy, *Classification, Clustering and Data Mining Applications*, D. Banks, L. House, F.R. McMorris, P. Arabie, W. Gaul (Eds), (Springer-Verlag, Berlin-Heidelberg, 2004), pp. 639–648.
28. I. Guyon and A. Elisseeff, An introduction to Variable and Feature Selection, *J. of Machine Learning Research*, Vol. 3 (MIT Press, Cambridge MA, 2003), pp. 1157–1182.
29. A.K. Jain, R.P.W. Duin and J. Mao, Statistical Pattern Recognition: A review, *Trans. on Pattern Analysis and Machine Intelligence*, vol. 22 (IEEE, USA, 2000), pp. 4–37.
30. C.H. Ooi and P. Tan, Genetic algorithms applied to multi-class prediction for the analysis of gene expression data, *Bioinformatics*, Vol. 19 (Oxford University Press,

Oxford, 2003), pp. 37–44.

31. I. Guyon, J. Weston, S. Barnhill and V. Vapnic, Gene Selection for Cancer Classification using Support Vector Machines, *Machine Learning*, Vol. 46 (2002), pp. 389–422.
32. B.L. Welch, The generalization of 'students' problem when several different population variances are involved, *Biometrika*, Vol. 34 (Oxford University Press, Oxford, 1947), pp. 28–35.
33. W. Pan, A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments, *Bioinformatics*, Vol. 18 (Oxford University Press, Oxford, 2002), pp. 546–554.
34. G.V. Trunk, A Problem of Dimensionality: A Simple Example, *Trans. on Pattern Analysis and Machine Intelligence*, Vol. 1 (IEEE, USA, 1979), pp. 306–307.
35. A.K. Jain and B. Chandrasekaran, Dimensionality and Sample Size Considerations in Pattern Recognition Practice, *Handbook of Statistics*, Vol 2 (Amsterdam: North-Holland,1982), pp. 835–855.
36. J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, V. Vapnik, Feature selection for SVMs, *In Advances in Neural Information Processing Systems* Vol. 13, 11th edition, Edited by Solla SA, Leen TK, Muller K-R. (MA: MIT press, Cambridge, 2001).
37. V. Vapnik, *Statistical Learning Theory*, (John Will and Sons, New York, 1998).
38. C. Burges, *A Tutorial on Support Vector Machines for Pattern Recognition*, (Kluwer Academic Publishers, Boston, 1998).
39. C.C. Chang and C.J. Lin, *LIBSVM : a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm
40. *Stanford Microarray Database*, http://genome-www5.stanford.edu, accessed 10 Jan. 2004.
41. M. Xiong, W. Li, J. Zhao, L. Jin and E. Boerwinkle, Feature (Gene) Selection in Gene Expression-Based Tumor Classification, *Mol. Genet. Metab.*, Vol. 73 (2001), pp. 239–247.
42. H. Zang, C.Y. Yu, B. Singer and M. Xiong, Recursive partitioning for tumor classification with gene expression microarray data, *Proceedings of the National Academy of Sciences*, Vol. 98 (National Academy of Sciences, USA, 2001), pp.6730–6735.
43. M.D. Radmacher, L.M. McShane and R. Simon, A Paradigm for Class Prediction Using Gene Expression Profiles, *J. of Computational Biology*, Vol. 9 (Mary Ann Liebert, Inc., New York, 2002), pp. 505–511.
44. D.M. Parkin, F.I. Braya and S.S. Devesa, Cancer burden in the year 2000. The global picture, *European Journal of Cancer*, Vol. 37 (Elsevier Science, 2001) pp. S4–S66.
45. W. Travis, T. Colby, B. Shimosato and E.Y.&Brambilla, *WHO International Histological Classification of Tumors: Histological Typing of Lung and Pleural Tumors*(Springer, Heidelberg, 1999).
46. M. Diehn, G. Sherlock, G. Binkley, H. Jin, J.C. Matese, T. Hernandez-Boussard, C.A. Rees, J.M. Cherry, D. Botstein, P.O. Brown and A.A. Alizadeh, SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data, *Nucleic Acids Research*, Vol. 31 (Oxford University Press, Oxford, 2003), pp. 219–223.
47. E. Alpaydin and C. Kaynak, Cascading Classifiers, *Kybernetika*, Vol. 34 (1998), pp. 369–374.
48. K. Pelckmans, J. De Brabanter, J.A.K Suykens and B. De Moor, Maximal Variation and Missing Values for Componentwise Support Vector Machines, *in Proc. of the International Joint Conference on Neural Networks*, (IJCNN 2005), Montreal, Canada, Aug. 2005 (to appear).

49. P.O'Neill, G.D. Magoulas and X. Liu, Improved Processing of Microarray Data Using Image Reconstruction Techniques, *Trans. on Nanobioscience*, Vol. 2 (IEEE, USA, 2003), pp. 176–183.