BMC Bioinformatics

BioMed Central

Research article

Regularized Least Squares Cancer Classifiers from DNA microarray data

Nicola Ancona*[1], Rosalia Maglietta[1], Annarita D'Addabbo[1], Sabino Liuni[2] and Graziano Pesole[2,3]

Address: [1]Istituto di Studi sui Sistemi Intelligenti per l'Automazione, CNR, Via Amendola 122/D-I, 70126 Bari, Italy, [2]Istituto di Tecnologie Biomediche-Sezione di Bari, CNR, Via Amendola 122/D, 70126 Bari Italy and [3]Dipartimento Scienze Biomolecolari e Biotecnologie, Universitá di Milano, Via Caloria 26, 20133 Milano, Italy

Email: Nicola Ancona* - ancona@ba.issia.cnr.it; Rosalia Maglietta - maglietta@ba.issia.cnr.it; Annarita D'Addabbo - daddabbo@ba.issia.cnr.it; Sabino Liuni - sabino.liuni@ba.itb.cnr.it; Graziano Pesole - graziano.pesole@unimi.it

* Corresponding author

from Italian Society of Bioinformatics (BITS): Annual Meeting 2005
Milan, Italy, 17–19 March 2005

Published: 1 December 2005

BMC Bioinformatics 2005, **6**(Suppl 4):S2 doi:10.1186/1471-2105-6-S4-S2

Abstract

Background: The advent of the technology of DNA microarrays constitutes an epochal change in the classification and discovery of different types of cancer because the information provided by DNA microarrays allows an approach to the problem of cancer analysis from a quantitative rather than qualitative point of view. Cancer classification requires well founded mathematical methods which are able to predict the status of new specimens with high significance levels starting from a limited number of data. In this paper we assess the performances of Regularized Least Squares (RLS) classifiers, originally proposed in regularization theory, by comparing them with Support Vector Machines (SVM), the state-of-the-art supervised learning technique for cancer classification by DNA microarray data. The performances of both approaches have been also investigated with respect to the number of selected genes and different gene selection strategies.

Results: We show that RLS classifiers have performances comparable to those of SVM classifiers as the Leave-One-Out (LOO) error evaluated on three different data sets shows. The main advantage of RLS machines is that for solving a classification problem they use a linear system of order equal to either the number of features or the number of training examples. Moreover, RLS machines allow to get an exact measure of the LOO error with just one training.

Conclusion: RLS classifiers are a valuable alternative to SVM classifiers for the problem of cancer classification by gene expression data, due to their simplicity and low computational complexity. Moreover, RLS classifiers show generalization ability comparable to the ones of SVM classifiers also in the case the classification of new specimens involves very few gene expression levels.

Background

The advent of the technology of DNA microarrays constitutes an epochal change in the study, treatment, analysis, classification and discovery of different types of cancer. It is well understood that cancer classification is a crucial step for cancer diagnosis and treatment [1,2]. Conventional classification of cancer has been based primarily on examination of the morphological appearance of tissue specimens, but this method suffers of serious limitations. It is subjective and depends on highly trained pathologists. Moreover, tumors with similar histopathological appearances can follow different clinical courses and show different responses to therapy [3]. The information provided by DNA microarrays allows to approach the problem of cancer diagnosis and treatment from a *quantitative* rather than *qualitative* point of view. The importance of the information embedded in gene expression data provided by DNA microarrays for identifying new cancer classes and for automatically classifying tumors to known classes was firstly pointed out by Golub in [1]. In tumor classification, the problem is to assign a label y, for example normal or cancerous tissue, to a new gene expression pattern $\mathbf{x}$, starting from the knowledge of ℓ *examples* $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2),...(\mathbf{x}\ell, y\ell)\}$ whose association between the gene expression pattern $\mathbf{x}_i$ and its relative class label y_i is known in advance. Here $\mathbf{x}$ is a vector whose components indicate the gene expression levels provided by a DNA microarray. Under this perspective, the problem of cancer classification can be seen as a supervised learning problem, or a *learning from examples* problem [4], in which the goal is to determine a separating surface, optimal under certain conditions, which is able to discriminate normal from cancer tissues, or to distinguish among different types of tumors. In this paper we focus only on two class classification problems, since multi-class problems can be seen as a straightforward generalization of two-class problems. Before introducing the main aspects of our work, it is worth to point out that the ultimate goal of any classifier, and in general of any learning machine, is *to generalize*, that is to predict the correct output y relative to never seen before input patterns $\mathbf{x}$, by using a training set S composed of a *finite* number of examples. Thus the central problem is not classifying the training data in S, because any sufficiently complex learning machine could separate S without errors. The crucial problem is to design classifiers having low error rate on new data. In the context of classification of DNA microarrays, such a problem is even more challenging because typically the number of examples is relatively small and the dimensionality, i.e. the number of genes whose expression levels are measured, is very large.

Statistical learning theory [5] provides a valuable non asymptotic theory for asking questions about the accuracy of models built when a limited amount of data is available. In this general framework, Support Vector Machine (SVM) classifiers provide excellent performances in terms of generalization error in different application domains such as object detection in images [6,7], odor classification [8], pedestrian detection [9], etc. In particular, in the context of cancer classification from gene expression data it outperforms many well known approaches [10-13] and it has to be considered as the method of reference for evaluating new techniques. The basic idea of statistical learning theory is very simple: for a finite set of training examples, the search for the best model or approximating function has to be constrained by an appropriately small hypothesis space, that is the set of functions the machine implements. If the space is too large, functions can be found which fit exactly the data, but they will have poor generalization capabilities on new data. SVM implements such an idea determining the classifier minimizing both the error on the training set (empirical risk) and the complexity of the hypothesis space.

Another approach to classification and in general to the problem of approximating a multivariate function from sparse data and in the presence of noise is regularization theory [14-16]. Also in this framework we need to constraint the hypothesis space for finding a suitable approximating function from a finite number of training examples. Such a constraint takes the form of a smoothness functional measuring the complexity of functions belonging to the chosen hypothesis space. In this general framework, Regularized Least-Squares (RLS) classifiers [17] provide a highly viable alternative to SVMs because they enjoy a number of suitable properties such as simplicity and reliability.

A first comparison between SVM and RLS classifiers can be found in [18]. In their analysis, the authors used very simple bench-mark data sets having characteristics very different from the ones relative to the cancer classification problem by gene expression data. In fact, they used data sets having a ratio between number of examples and number of components ranging from 3.5 for the sonar data set to 96 for the pima indian data set. Such ratios are very far from the ratio of order of 1/100 that is typical for the problem we are considering here. So from their study we can not infer any consequence about the performances of the RLS classifiers on the problem at hand. In this paper we compare SVM and RLS classifiers for the specific problem of cancer classification by gene expression data. In the context of supervised learning models, as the ones we are considering here, particularly important is the quantity to measure for comparing two machines. We know that two machines have similar performances if their generalization errors are comparable. As we will show in the next sections [5], a measure of the generalization error of any supervised learning machine is the *risk* and so models showing the same risk have comparable performances. However, the risk functional, as usually defined, has not a practical usefulness because it involves the knowledge of the probability distribution function underlying the data that is in general unknown. Nevertheless, we can adopt the Leave-One-Out (LOO) procedure which uses the available data for evaluating the generalization error of a machine. In fact, as the Luntz and Brailovsky theorem shows [19], the LOO error is an almost unbiased estimator of the risk and so it is a practical procedure for assessing the performances of a supervised learning machine from a finite number of data. Based on this estimator we show that RLS and SVM models have similar generalization abilities. The comparison involved three different data sets described in [1,12,20]. The experimental results suggest that we can benefit of the simplicity of RLS machines maintaining the same prediction error of SVM. The main advantage of RLS machines is that for solving a classification problem we need to solve a single linear system of order equal to either the number of features or the number of training examples. This is in contrast to SVM approach which requires the solution of a quadratic programming problem with linear constraints. Moreover and more important, RLS machines allow to get an exact measure of the leave-one-out error with just one training. In the case of SVM, such important measure requires the training of a number of machines equal to the number of training examples. At the aim of fully assessing both the classification models, we analyze their performances with respect to the number of genes, selected with different gene selection strategies. Note that the focus here is not on the determination of the optimal number of genes for classifying tissues belonging to a given tumor class. For this reason others and more sophisticated methodologies have to be adopted which take into account the bias selection problem [21]. Here we want to show that both models have comparable performances even when a very few number of genes is used for classifying. Following the statistical approach outlined by Golub and its co-workers in [1,2,12], we adopt non-parametric permutation tests for studying how many and what genes have to be used for classifying. The problem of identifying the gene signature concerning a particular type of cancer is out of the scope of the present work.

Methods

Classification models

Due to the particular problem of cancer classification from gene expression data in which we have a small number of training examples, each one with very large dimensionality, we limit our attention to linear classifiers. Nevertheless the methods we are going to illustrate can be easily generalized for designing non-linear classifiers.

We are given a training set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell}$ of size ℓ where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \{-1,1\}$, for $i = 1,2,\ldots,\ell$. In the simplest case of linearly separable set S, the classification problem consists of determining a hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$, where $\mathbf{w} \in \mathbb{R}^n$ and $b \in \mathbb{R}$, such that: $y_i = \text{sign}(\mathbf{w} \cdot \mathbf{x}_i + b)$ for $i = 1,2,\ldots,\ell$, where sign (x) is 1 if $x \geq 0$ and -1 otherwise. Actually, classification is an ill-posed problem [14] because an infinite number of solution exist and then some constraint has to be imposed to the problem for making the solution unique.

SVM classification

The constraint imposed by SVM on the classification problem is the following: the solution has to maximize the distances with the closest points of S. The optimal separating hyperplane found by SVM, in the case of linearly separable set S, is the one maximizing the margin m, where m = $2/||\mathbf{w}||$ is the distance between the hyperplane and the closest points of S. In the general hypothesis of non linearly separable classes, the optimal separating hyperplane $\mathbf{w}^* \cdot \mathbf{x} + b^* = 0$ found by SVM is solution of the following quadratic programming (QP) problem P1 with linear constraints:

$$\begin{array}{ll} \max\limits_{\boldsymbol{\lambda}} & -\frac{1}{2}\boldsymbol{\lambda}^T \boldsymbol{D}\boldsymbol{\lambda} + \sum_{i=1}^{\ell} \lambda_i \\ \textit{subject to} & \sum_{i=1}^{\ell} \lambda_i y_i = 0 \\ & 0 \leq \lambda_i \leq C \quad i = 1,2,\ldots,\ell \end{array}$$

where $\boldsymbol{D}$ is a matrix of size $\ell \times \ell$, with $D_{ij} = y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j$ for i,j = 1, 2, ...,ℓ and $\lambda = (\lambda_1, \lambda_2,\ldots, \lambda\ell)^T$ is a vector of ℓ non negative Lagrange multipliers. The regularization parameter C is the only free parameter and its value can be chosen by using cross validation. Let λ^* be the solution of the considered problem P1. So the optimal $\mathbf{w}^*$ is:

$$\mathbf{w}^* = \sum_{i=1}^{\ell} \lambda_i^* y_i \mathbf{x}_i \qquad (1)$$

and the optimal b^* is given by:

$$b^* = y_i - \mathbf{w}^* \cdot \mathbf{x}_i \qquad (2)$$

for each i such that $0 < \lambda_i^* < C$. The points $\mathbf{x}_i$ with $\lambda_i^* > 0$ live on the margin of the classes and they are called *support vectors*. The classification of a new data $\mathbf{x}$ involves the evaluation of the decision function:

$$y = \text{sign}(f(\mathbf{x})) \qquad (3)$$

where:

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} \lambda_i^* y_i \mathbf{x}_i \cdot \mathbf{x} + b^*$$

RLS classification

RLS models [14] were proposed mainly for facing regression problems. The main difference between a regression and classification problem is that in the former the output variable y can assume any real value; in the latter, it can assume a finite number of possible values. In our case, y assumes only two values {-1,1}. This means that every classification problem can be considered as a regression problem. In the case of linear regression, we want to determine the function f($\mathbf{x}$) = $\mathbf{w} \cdot \mathbf{x}$, with $\mathbf{w} \in \mathbb{R}^n$, which approximates the examples in S in the least squares sense. This is equivalent to solving the following constrained minimization problem P2:

$$\min_{\mathbf{w}} \quad \frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - \mathbf{w}\cdot\mathbf{x}_i)^2$$

subject to $||\mathbf{w}||^2 \leq \alpha$

where $\alpha \in \mathbb{R}$ and $||\mathbf{w}|| = \sqrt{\mathbf{w}\cdot\mathbf{w}}$ is the Euclidean norm induced by the scalar product $\cdot$. Note that the bias term is implicitly present in our model by including a component constant and equal to one to the input vectors. Before solving the problem P2, some considerations are in order. The objective function that we minimize, in this particular case, takes the form of the mean square error of the predictor $y = \mathbf{w}\cdot\mathbf{x}$ evaluated on the training data. Here the error is expressed as the square *deviation*, $\varepsilon_i = (y_i - \mathbf{w}\cdot\mathbf{x}_i)^2$, between the target value y_i and the value of the predictor $\mathbf{w}\cdot\mathbf{x}_i$. Let d_i be the square *distance* between the generic input data $\mathbf{x}_i$ and the approximating hyperplane $y = \mathbf{w}\cdot\mathbf{x}$, where by definition:

$$d_i = \frac{\varepsilon_i}{1+\|\mathbf{w}\|^2} \qquad (4)$$

This equation shows that the smaller $||\mathbf{w}||^2$, the better the deviation ε_i approximates the true distance d_i. This is the reason why we introduce the constraint $||\mathbf{w}||^2 \leq \alpha$. In this way the optimal approximating hyperplane solution of the constrained problem is the hyperplane minimizing the mean square distance with the training points. For determining $\mathbf{w} \in \mathbb{R}^n$ solution of P2, let us consider the Lagrangian function:

$$L(\mathbf{w}) = \frac{1}{l}\sum_{i=1}^{\ell}(y_i - \mathbf{w}\cdot\mathbf{x}_i)^2 + \lambda(\|\mathbf{w}\|^2 - \alpha) \qquad (5)$$

The vector $\mathbf{w}$ minimizing (5) is solution of the following linear system of order *n*:

$$(\mathbf{X}\mathbf{X}^\top + \lambda\ell\mathbf{I}_n)\ \mathbf{w} = \mathbf{X}\mathbf{y} \qquad (6)$$

where $\mathbf{X}$ is a $n \times \ell$ matrix having the examples $\mathbf{x}_i$ as its columns, $\mathbf{y} = (y_1, y_2, \ldots, y\ell)^\top$ and $\mathbf{I}_n$ is the $n \times n$ identity matrix. Note that, since the matrix $\mathbf{X}\mathbf{X}^\top$ is positive semidefinite, then for $\lambda > 0$ the matrix $\mathbf{X}\mathbf{X}^\top + \lambda\ell\mathbf{I}_n$ is definite positive and therefore invertible. Then the vector $\mathbf{w}^*$ solution of the problem P2 exists and it is given by:

$$\mathbf{w}^* = (\mathbf{X}\mathbf{X}^\top + \lambda\ell\mathbf{I}_n)^{-1}\mathbf{X}\mathbf{y} \qquad (7)$$

It is possible to show that the value of λ controls the influence of the noise present in the data on the estimation of the solution $\mathbf{w}^*$. The parameter λ, called regularization parameter, is the only free parameter and its value can be chosen by using cross validation. Analogously to SVM, the classification of a new data $\mathbf{x}$ involves the evaluation of the decision function:

$$y = \text{sign}(\mathbf{w}^* \cdot \mathbf{x}) \qquad (8)$$

As equation (7) shows, determining $\mathbf{w}^*$ requires the solution of a linear system of *n* order, where *n* is the number of components of each $\mathbf{x}_i$. In some cases *n* could be extremely large and so any direct method can be adopted for estimating $\mathbf{w}^*$. This occurs in the problem at hand where the number of genes *n* of each specimen is order of tens of thousand and the number ℓ of specimens is order of ten or hundred. We will show that the models we are describing allow to rewrite a linear system of *n* order as a linear system of ℓ order, overcoming the difficulties connected to problems with a huge number of features. At this aim, let us suppose $\mathbf{w}$ to be expressed as linear combination of the vectors $\mathbf{x}_i$ for $i = 1,2, \ldots,\ell$. This means that there exist ℓ coefficients $\mathbf{c} = (c_1, c_2, \ldots, c\ell)^\top$ such that:

$$\mathbf{w} = \mathbf{X}\mathbf{c} \qquad (9)$$

Substituting (9) in (6) we have:

$$(\mathbf{K} + \lambda\ell\mathbf{I}\ell)\ \mathbf{c} = \mathbf{y} \qquad (10)$$

where $\mathbf{K} = \mathbf{X}^\top\mathbf{X}$ is a $\ell \times \ell$ matrix with generic element $\mathbf{K}_{ij} = \mathbf{x}_i \cdot \mathbf{x}_j$ and $\mathbf{I}\ell$ is the identity matrix of ℓ order. Also in this case, since $\mathbf{K}$ is a positive semidefinite matrix, then for $\lambda > 0$ the matrix $\mathbf{K} + \lambda\ell\mathbf{I}\ell$ is positive definite and so invertible. Then the vector $\mathbf{c}^* \in \mathbb{R}^l$ solution of (10) is given by:

$$\mathbf{c}^* = (\mathbf{K} + \lambda\ell\mathbf{I}\ell)^{-1}\mathbf{y} \qquad (11)$$

obtained by solving a linear system of ℓ order. Note that the normal $\mathbf{w}^*$ to the optimal approximating hyperplane can be recovered by using (9). In this case the classification of a new data $\mathbf{x}$ involves the evaluation of the decision function:

$$y = \text{sign}\left(\sum_{i=1}^{\ell} c_i^* \mathbf{x}_i \cdot \mathbf{x} \right) \qquad (12)$$

Comparison between SVM and RLS classifiers

Numerous differences exist between these two classification models, but we will only mention some of these which are relevant for our discussion. The first difference consists in the method employed for determining the optimal **w**. SVM requires the solution of a QP problem with linear constraints of order ℓ, while RLS requires the solution of a system of linear equation of order ℓ or n. In the former, the complexity in solving problem P1 is independent of n. Moreover, when the number ℓ of examples is extremely large, decomposition methods can be applied for determining the exact solution [22]. In the latter, the complexity depends on ℓ and n. When both these quantities assume large values, iterative schemes have to be adopted for solving the system (6) or (10), so providing only approximated solutions.

The second difference consists in the representation of the optimal **w**. In SVM (see equation (1)), the solution is *sparse* meaning that it is expressed as linear combination of a fraction of the training examples (support vectors). In RLS (see equation (9)), on the contrary, the solution is *dense* meaning that it is expressed as linear combination of all training examples.

Leave-one-out error

As we have already said in the introduction, the ultimate goal of a supervised learning machine y = f (x, α) is to generalize, that is to correctly predict the output y corresponding to never seen before input patterns x. Here α is a parameter vector which the machine depends on, for example C in SVM and λ in RLS classifiers. Then a comparison between different classification models has to involve the comparison of their generalization errors. A measure of the generalization error of such a machine f is the risk R[f] defined as the expected value of the loss function V(y,f(x, α)) (see [5]):

$$R[f] = \int V(y, f(\mathbf{x}, \alpha)) p(\mathbf{x}, y) d\mathbf{x} dy \qquad (13)$$

where $p(\mathbf{x}, y)$ is the probability density function underlying the data. The particular form of the loss function depends on the problem at hand. In classification problems, the loss takes the form of:

$$V(y, f(\mathbf{x}, \boldsymbol{\alpha})) = \begin{cases} 0 & \text{if } y = f(\mathbf{x}, \boldsymbol{\alpha}) \\ 1 & \text{if } y \neq f(\mathbf{x}, \boldsymbol{\alpha}) \end{cases} \qquad (14)$$

In general the probability density function $p(\mathbf{x}, y)$ is unknown and so we are not able to evaluate the risk. The only data we have are ℓ observations (examples) $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell}$ of the random variables **x** and y drawn according to $p(\mathbf{x}, y)$. The leave-one-out (LOO) error provides a measure of the generalization error of a learning machine by using the ℓ observations in S. In fact, as the Luntz and Brailovsky theorem shows [19], the LOO error is an almost unbiased estimator of the risk (13) and it allows of assessing the performances of a supervised learning machine from a finite number of data. The computation of the LOO error is very simple. For every i = 1, 2, ..., ℓ, let f_{S^i} be the machine trained on the set $S^i = \{(\mathbf{x}_1, y_1),...,(\mathbf{x}_{i-1}, y_{i-1}), (\mathbf{x}_{i+1}, y_{i+1}),...,(\mathbf{x}\ell, y\ell)\}$ obtained from S removing the i-th example. Test the function f_{S^i} on the left out example $(\mathbf{x}_i, y_i)$ and measure the value of the loss function $V(y_i, f_{S^i}(\mathbf{x}_i, \alpha))$. Repeat this procedure for each of the ℓ examples of the training set and sum the errors made. This number is the LOO error:

$$\mathcal{L}(s) = \sum_{i=1}^{\ell} V(y_i, f_{S^i}(\mathbf{x}_i, \boldsymbol{\alpha})) \qquad (15)$$

Note that $\mathcal{L}(s)$ is the quantity that we have to compute for measuring the performances of any supervised learning machine, because it provides an estimate of the risk or generalization error associated to the selected machine. Moreover, this is the procedure of choice for estimating the unknown parameter vector α which the machine depends on. In fact, for a fixed training set, the generalization error of the machine is a function of α. Then, the best parameter vector α^* will be the one minimizing $\mathcal{L}(s)$.

LOO-error for RLS classifiers

Although the LOO error enjoys several interesting properties, its computation is tremendously expensive because it requires of training a number of machines equal to the number of training examples. In the case of RLS classifiers, the LOO error can be calculated in an exact way just training a single machine by using all the training examples. In fact it can be showed [15] that:

$$\mathcal{L}(s) = \sum_{i=1}^{\ell} V\left(y_i, sign\left(y_i - \frac{y_i - f_S(\mathbf{x}_i, \boldsymbol{\alpha})}{1 - (\mathbf{KG})_{ii}} \right) \right) \qquad (16)$$

where f_s is the machine trained on S and $\mathbf{G} = (\mathbf{K} + \lambda \ell \mathbf{I} \ell)^{-1}$. This is a fundamental property of the RLS classifiers because it allows to evaluate the generalization ability of a classifier without any additional cost.

Gene selection

A very important question in cancer classification problem is determining which genes are the most relevant in identifying a specimen or a particular disease. This is an open problem, relevant for several reasons both biological and computational. Finding genes which expression levels correlate with a particular disease is important for understanding the disease and for choosing the most appropriate treatment. Furthermore, classifying a specimen on the basis on few expression levels could in principle improve the performances of the classifier, eliminating the noise associated to irrelevant genes. Gene selection is a particular instance of a more general problem known in machine learning as feature selection. In general, the methods for selecting features can be grouped in two main categories: filter methods and wrapper methods [23]. Filter methods select features by using criteria independent of the ones used in the classification stage. Wrapper methods, on the contrary, use the same or similar criteria as the ones used by the classifier. In this paper we focus on two different feature selection approaches. The first one [1,2], known as signal-to-noise (S2N), is a filter method and it is based on the following statistic:

$$T_{S2N}(j) = \frac{\mu_{+1}(j) - \mu_{-1}(j)}{\sigma_{+1}(j) + \sigma_{-1}(j)} \quad j = 1, 2, ..., n \qquad (17)$$

where *j* is the gene index. $(\mu_{+1}(j), \sigma_{+1}(j))$ and $(\mu_{-1}(j), \sigma_{-1}(j))$ are the mean and the standard deviation of the expression levels of the *j*-th gene in the positive and negative examples respectively. Genes x_j highly correlated with the class label or more relevant for classifying are expected to provide large values of $|T_{S2N}(j)|$ The second approach we consider is a variant of *recursive feature elimination* (RFE) strategy proposed in [24]. It is a wrapper method and it is based on the following statistic:

$$T_w(j) = w_j \quad j = 1,2,...,n \qquad (18)$$

where **w** is the normal of the optimal separating hyperplane found by SVM or RLS methods. The idea underlying this approach is very simple. We know that the label *y* associated to a new input **x** is given by $y = \text{sign}\left(\sum_{j=1}^{n} w_j x_j + b\right)$. So, if the gene expression levels have similar ranges, genes having large values of $|w_j|$ are more important than others in determining the class label. Instead of using a recursive approach for selecting the most relevant genes as suggested in [24], we use a more greedy strategy consisting in training the machine one time only by using all the available genes and selecting the most informative features according to the obtained **w**. For this reason we call our approach *not-RFE* (NRFE). In both strategies, the genes are ranked in decreasing order according to the selected statistic and the highest values correspond to the most relevant genes.

Number of relevant genes

So far we have described two statistics for ranking genes based on their expression levels in both classes. Now, in order to determine how many genes are really important for classifying a given specimen, we apply a common method in classical statistics named hypothesis testing (see [1]). The idea is to hypothesize that there is no dependency between expression levels and class labels, and to consider relevant for the classification those genes which reject such hypothesis. At this aim, we define the null hypothesis H_0 in which we assume that the random variables x and y are independent or equivalently that the class conditional probability density functions are identical. The goal of hypothesis test is to reject H_0 at a given level of significance α, where α is the probability of rejecting the null hypothesis when it is true, that is of declaring that the x and y are uncorrelated when they are not. Let t_0 be the observed value of the statistic T as computed on the data set S, $t_0 = T(x_1, y_1, x_2, y_2,...,x\ell, y\ell)$, and let $p_0 = P_T(T \geq t_0)$ be the corresponding p-value, that is the probability that T is grater than or equal to t_0. Note that P_T is the distribution function of the random variable T under the null hypothesis. If $p_0 \leq \alpha$ then we reject H_0 at level of significance α.

The application of the hypothesis testing method requires the knowledge of the density or distribution function of the adopted *T* statistic under the null hypothesis. When the density of the adopted statistic is unknown or when the data do not verify the hypotheses which the statistic is based on, then we have to the invoke nonparametric permutation tests [25]. This nonparametric technique allows to estimate the probability density function of any statistic, under the null hypothesis, from the available data. The reason which justifies this procedure for estimating the density $p_T(t)$ is that under the null hypothesis, since the random variables *x* and *y* are independent, all the training set generated through permutations are equally likely to be observed.

Results

Data sets description

The above mentioned classification techniques have been applied to different cancer diagnosis problems. Three benchmark data sets have been considered. The first one, named 'Leukemia data set' [1], concerns the classification of acute leukemias into acute myeloid leukemia (AML) and acute lymphoblastic one (ALL). It consists of 38 bone marrow samples (11 AML, 27 ALL) obtained from acute leukemia patients at the time of diagnosis (i.e. before that any treatment was used). These samples are used as training set. An additional set, composed of 14 AML and 20

ALL samples, is utilized to test the classifiers. Each sample is a vector composed by 7129 elements, each one corresponding to the $\log_{10}$ normalized expression value of a gene. This data set has been extensively analyzed in literature [2] also by using machine learning techniques [10]. Much more details about this data set and a complete breakdown of microarray composition can be found on the web site http://www.genome.wi.mit.edu/MPR. The second data, named 'Colon data set' [20], regards the problem of classifying tumor and normal colon tissues. It is composed by 40 tumor and 22 normal colon tissue samples. Each sample consists of 2000 human gene expression levels. The data set and more detailed information on it are available on the web at site http://www.molbio.princeton.edu/colondata. The last analyzed data set is relative to the classification of different malignancies samples against normal tissue ones [12], it will be identified as 'Multi-cancer data set'. It is composed by 280 samples: 190 examples are relative to cancer tissues, spanning 14 common tumor types, the remaining 90 samples represent normal tissues. Each example in this data set consists of the expression levels of 16063 genes. Complete details regarding patient samples, pathology, molecular biology protocols, data analysis and additional information are available at site http://www.genome.wi.mit.edu/MPR/GCM.html. It is worth nothing that, in the present work, this data set is analyzed in order to perform a two-class classification problem (i.e. to discriminate between diseased and normal samples).

Results on the Leukemia data set

First of all, we used all of 7129 gene expression levels present in each specimen. We trained SVM classifiers on the 38 samples in the training set for different values of C parameter, measuring for each one the empirical risk and the LOO error given by equation (15). The training set is linearly separable and the LOO error reaches its minimum value of 1 (see table 1) in correspondence of C = 1e - 6. Then the best SVM classifier on this training set is the one obtained with C = 1e - 6 because it is the machine minimizing the LOO error. We tested such machine on the 34 points in the test set obtaining 1 error (see table 1). The same results are reported in [10], where the authors also noted that using SVM with polynomial kernel functions did not improve the performances.

The same procedure was carried out by using RLS machines. We trained RLS classifiers on the training set for different values of λ parameter and for each one we measured the empirical risk and the LOO error by using equation (16). The training set is linearly separable for each λ in the considered range. Moreover, the LOO error reaches its minimum value of 1 (see table 1) in correspondence of $\lambda = 1$. Then the best RLS classifier on this training set is the one obtained with $\lambda = 1$. We tested such machine on the test set obtaining 1 error as reported in table 1.

Table 1: Minimum LOO error on the Leukemia training set (composed of 38 examples), error on the test set (34 examples) and minimum LOO error on the whole Leukemia data set (72 examples).

	SVM	RLS
LOO error on training set	1	1
Test error	1	1
LOO error on the whole data set	1	1

Successively, in order to carry out a most accurate analysis and to have a most complete insight about the performances of SVM and RLS machines on this data set, we have computed the LOO errors on the whole data set obtained putting together training and test examples. The values of the free parameters, corresponding to the best machines, are $C = 45$ and $\lambda = 10$ respectively. The LOO errors obtained by the best machines are reported in table 1 where clearly results that SVM and RLS behave exactly the same. In order to understand the influence of irrelevant genes on the performances of the classifiers, we considered some subsets of features. We established the number of genes to select applying permutation tests to the data, by using T_{S2N} and T_w statistics. Figure 1 depicts the values of T_{S2N} statistic as computed on the actual data set and on randomly permuted class labels. The number of permutations of the labels was 1500. Genes more highly expressed in ALL are shown in the left picture, and those more highly expressed in AML are shown in the right picture. The large number of genes highly correlated with the class distinction is clear from the picture. Moreover, in both pictures, the curve of the observed statistic intersects the 5% curve about at 1000 genes, indicating that in the data set there are 1000 genes which reject H_0 at significance level of $\alpha = 5\%$. Then we ranked the genes according to the absolute value of $T_{S2N}$and chose the top k genes, with k equal to 1000, 100, 50, 40, 30, 20, 10, 5 and 3 genes. A similar analysis has been effectuated by using T_w statistic. Here $\mathbf{w}$ was the parameter vector corresponding to the best RLS classifier, that is the one minimizing the LOO error on the current data set[1]. As picture 2 shows, T_w is unable to disclose the correlation existing between gene expression level and class label, as T_{S2N} does. Nevertheless, we equally measured the performances of SVM and RLS classifiers on genes selected by T_w statistic. In fact, as noted in [2], in some particular cases, some genes may be truly predictive of the class label despite the lack of statistical significance in permutation tests. At the aim of testing this experimental evidence, we ranked the genes according to

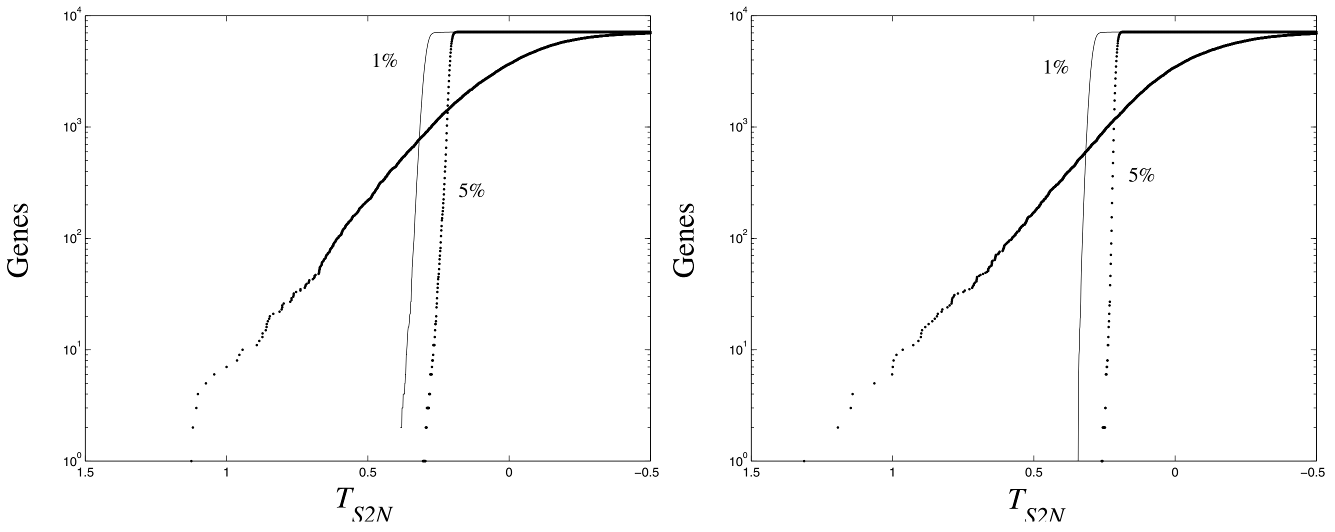

Figure 1
: Observed $T_{s2N}(j)$ distribution computed on the Leukemia data set, compared to randomly permuted class distinctions. The number of genes highly expressed in a) ALL and b) AML is shown on y-axis.

the absolute value of T_w and chose the top k genes, with k equal to 1000, 100, 50, 40, 30, 20, 10, 5 and 3 genes. In table 2 we report the performances of SVM and RLS classifiers obtained on the Leukemia data set, for various number of genes selected by S2N and NRFE methods.

Results on the Colon data set

First of all note that in this case, as in the following data set, we do not have the distinction between training and test set, because we have a single data set. For this reason, we do not report the test error but the LOO error only. We have primarily evaluated the performances of SVM and RLS classifiers on the Colon data set by using all the gene expression levels present in each specimen, successively we have consider opportune subsets of genes. The experimental results on the whole and reduced data sets are summarized in table 3.

Table 2: Minimum LOO error computed on Leukemia data set (composed of 72 examples), for various number of genes, selected with S2N and NRFE statistics.

	SVM		RLS	
genes	S2N	NRFE	S2N	NRFE
1000	1	1	2	1
100	1	0	1	0
50	1	0	2	0
40	2	0	2	0
30	2	0	2	0
20	2	0	2	0
10	2	1	2	0
5	1	1	2	2
3	4	3	4	2

The behavior of the empirical risk and of the LOO error of SVM and RLS classifiers evaluated for different values of the regularization parameter are depicted in figure 3 for the whole Colon data set. Note that the data set is linearly separable. These plots give also a precious hint to fully understand the role of free parameters (C in SVM and λ in the RLS machines) by observing the empirical risk curves. In fact, increasing C in SVM the empirical risk decreases, whereas increasing λ in RLS the empirical risk increases. These behaviors of the empirical risk curves can be fully justified reminding that, in SVM, the C parameter can be thought of as the cost the machine pays for each training error. On the contrary, in the RLS machines, the same role is played by $\frac{1}{\lambda}$ (see equation 5).

In order to determine the number of relevant genes to be considered in the feature selection process, we have computed the T_{S2N} statistic on the actual data set and in the hypothesis that H_0 holds true. The number of label permutations was 2000. The observed statistic intersects the 5% curve in correspondence of 500. So 500 is the maximum number of genes which reject the null hypothesis at significance level of 5%. Then we ranked the genes according to the absolute value of T_{S2N} and chose the top k genes, with k equal to 500, 400, 300, 200, 100, 50, 10 and 5 genes. A similar analysis has been effectuated by using T_w statistic. Also in this case, this statistic shows poor capacity of revealing the existing correlation in the data. As in the previous analysis, we ranked the genes according to the

Table 3: Minimum LOO error computed on Colon data set (composed of 62 examples), for various number of genes, selected with S2N and NRFE statistics.

	SVM		RLS	
genes	S2N	NRFE	S2N	NRFE
2000	6		6	
500	6	6	7	6
400	7	6	6	6
300	6	6	6	6
200	7	6	7	4
100	9	5	7	4
50	8	4	6	1
10	6	6	7	5
5	7	8	7	8

Table 4: Minimum LOO error computed on Multi-cancer data set (composed of 280 examples), for various number of genes, selected with S2N and NRFE statistics.

	SVM		RLS	
genes	S2N	NRFE	S2N	NRFE
16063	105		90	
1400	46	40	59	49
1000	42	41	57	50
500	50	41	57	56
300	51	38	57	54
200	51	50	55	50
100	63	97	51	58
50	59	76	43	61
10	63	74	59	73

absolute value of T_w and chose the top k genes, with k equal to 500, 400, 300, 200, 100, 50, 10 and 5 genes.

Results on the Multi-cancer data set

Primarily, the 16063 gene expression levels of each specimen have been used for classifying. The experimental results are summarized in table 4.

The data set is linearly separable. The best SVM corresponds to $C = 3$. The best RLS classifier corresponds to $\lambda = 20$.

It is important to note that the errors obtained on this data set are much greater than the one achieved in the data sets previously analyzed, probably reflecting the large complexity of the data due to the great degree of biological variability in gene expressions.

The non parametric permutation test was carried on the Multi-cancer data set, performing 1000 random permutations of the class labels. The maximum number of genes which rejects the null hypothesis at significance level of 5% is 1400. Then we ranked the genes according to the absolute value of T_{S2N} and chose the top k genes, with k equal to 1400, 1000, 500, 300, 200, 100, 50 and 10 genes. The same numbers of genes were selected by using the T_w statistic. The results on the reduced data sets are reported in table 4.

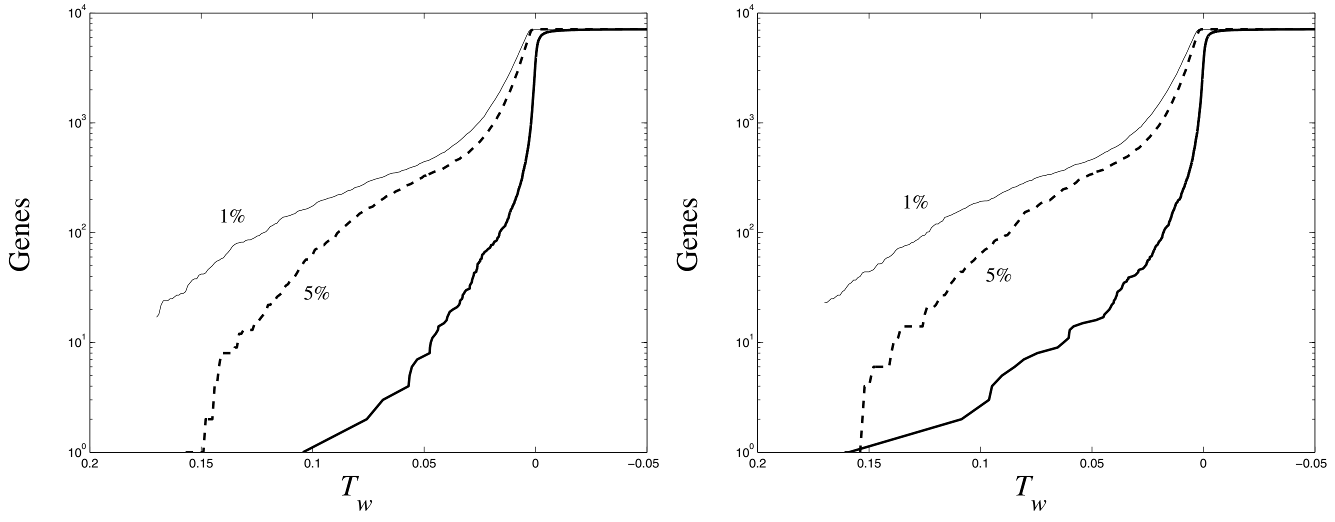

Figure 2

: Observed *Tw(j)* distribution computed on the Leukemia data set, compared to randomly per-mutated class distinctions. The number of genes highly expressed in a) ALL and b) AML is shown on y-axis.

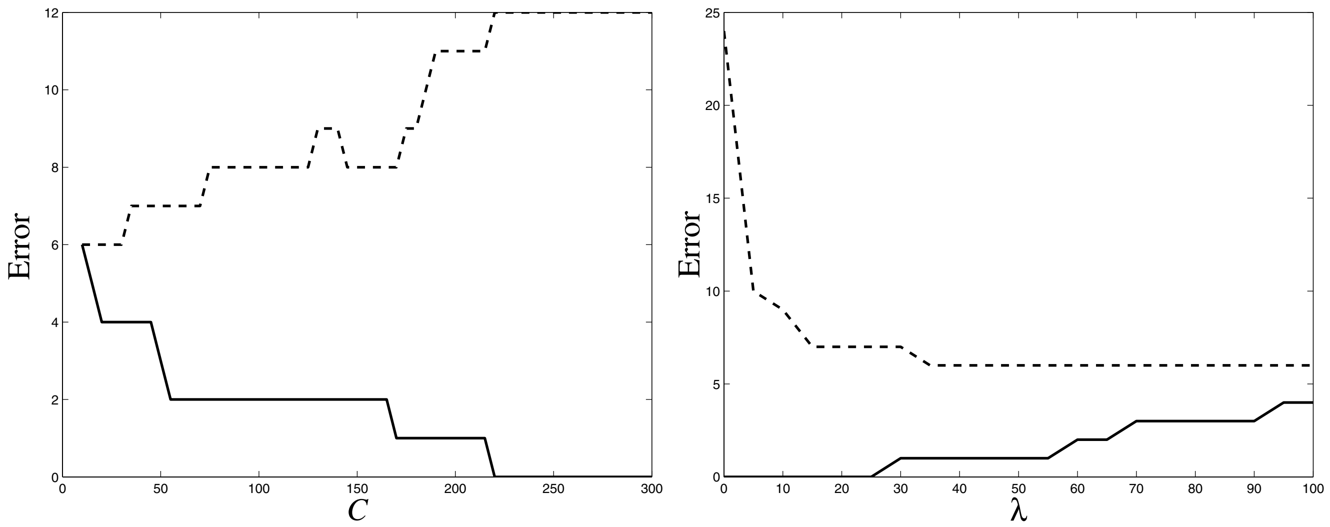

Figure 3
: LOO error (dotted line) and empirical risk (solid line) w.r.t the regularization parameter obtained on Colon data set by using a) SVM and b) RLS classifiers.

Discussion

Some conclusions on the two classification algorithms can be drawn. The first and more important one is that, when the whole data sets are considered, both machines provide generalization errors comparable as the tables 1, 3 and 4 show. This indicates that RLS approach is able to determine classifiers with good generalization ability even in the case of very small training set, with a huge number of features.

Concerning the computational time, both techniques require a few seconds for determining the optimal classifier because, in the present context, the training involves only a few examples. The second consideration concerns the role of λ in RLS machines. This parameter de-facto controls the generalization ability of the RLS classifiers, exactly as C does in SVM ones. The figure 3 depicting the behavior of the LOO error shows this fact. We have observed similar behaviors of this quantity in all the experiments carried out which do not show for lack of space. Moreover, our analysis shows that standard least-square machines, obtained setting $\lambda = 0$, have very poor generalization abilities. In fact for $\lambda = 0$ all the considered RLS classifiers separate correctly the training data, but they show a very large LOO error. The main problem in machine learning is not to correctly classify the training data. The main problem is to generalize and RLS classifiers guarantee high generalization ability for appropriate values of the regularization parameter λ.

The performances of SVM and RLS classifiers continue to be comparable even though the number of genes used for classifying a specimen is extremely reduced. Tables 2, 3 and 4 confirm such a result. Moreover, as noted in all three data set, also a statistic which is not able to reveal statistically significant differences in the data can however select genes which increase the performances of the classifier. This is not surprising. The fact that a gene is relevant for classifying a given specimen does not involve the statistic, it involves the classification process. So, a gene is relevant for a classifier if its usage reduces the *generalization error* of the classifier, as measured by the LOO error. Any gene selection strategy has to guarantee that the subset of genes selected is the most appropriate for the chosen classifier, that is it is the subset of features minimizing the LOO error of the classifier. In this sense feature selection and parameter selection are two instances of the same problem which has as ultimate objective the one of reducing the generalization error of any learning machine.

Conclusion

In this paper we have shown that RLS classifiers have performances comparable to the ones of SVM classifiers for the problem of cancer classification by gene expression data. The comparison has been carried on measuring the Leave-One-Out errors relative to each classifier obtained on three different real data set. The classification performance analysis involved the whole set of genes as well as suitable subsets of genes selected by different gene selection strategies. Our analysis suggests that RLS classifiers are a valuable alternative to SVM classifiers for the problem at hand due to their simplicity and low computational cost. Moreover, RLS classifiers show generalization errors comparable to the ones of SVM classifiers also in the case the classification of new specimens involve very few gene expression levels.

List of abbreviations used

SVM: Support Vector Machines. RLS: Regularized Least Square. LOO: Leave One Out.

Note

[1] Note that the T_w statistic can be considered a filter method when the features it selects are input to SVM classifiers, and a wrapper method when the features are input to RLS classifiers.

Acknowledgements

We would like to thank Sebastiano Stramaglia for valuable and illuminating discussions on numerous theoretic and experimental aspects of the paper. This work was supported by Cluster C03 *Studio di geni di interesse biomedico e agroalimentare*, FIRB project Bioinformatica per la Genomica e la Proteomica, PON avviso n.68, Contributi Straordinari del D.D. n. 1105 (Ministero dell Istruzione e Ricerca Scientifica, Italy) and AIRC.

References

1. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring". *Science* 1999, 286:531-537.
2. Slonim DK, Tamayo P, Mesirov JP, Golub TR, Lander ES: **Class Prediction and Discovery Using Gene Expression Data.** *Proceedings of the Fourth Annual Conference on Computational Molecular Biology (RECOMB)* 2000:263-272.
3. Tibshirani R, Hastie T, Narasimhan B, Chu G: **Diagnosis of multiple cancer types by shrunken centroids of gene expression.** *PNAS* 2002, **99:**6567-6572.
4. Poggio T, Girosi F: **A Theory of Networks for Approximation and Learning.** In *Tech Rep A.I. Memo No. 1140 Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA*; 1989.
5. Vapnik V: *Statistical Learning Theory John Wiley & Sons, INC*; 1998.
6. Heisele B, Poggio T, Pontil M: **Face Detection in Still Gray Images.** In *Tech Rep A.I. Memo No. 1687 Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA*; 2000.
7. Ancona N, Cicirelli G, Stella E, Distante A: **Ball detection in static images with Support Vector Machines for classification.** *Image and Vision Computing Elsevier* 2003, **21(8):**675-692.
8. Distante C, Ancona N, Siciliano P: **Support Vector Machines for olfactory signals recognition.** *Sensors and Actuators B Elsevier* 2003, **88(1):**30-39.
9. Papageorgiou C, Evgeniou T, Poggio T: **A trainable pedestrian detection system.** In *Proceedings of Intelligent Vehicles Stuttgart, Germany*; 1998.
10. Mukherjee S, Tamayo P, Mesirov JP, Slonim D, Verri A, Poggio T: **Support Vector Machine Classification of Microarray Data.** In *Tech Rep CBCL Paper #182/AI Memo #1677 Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA*; 1999.
11. Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D: **Knowledge-based analysis of microarray gene expression data by using support vector machines.** *PNAS* 2000, **97:**262-267.
12. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR: **Multiclass cancer diagnosis using tumor gene expression signatures.** *PNAS* 2001, **98:**15149-15154.
13. Romualdi C, Campanaro S, Campagna D, Celegato B, Cannata N, Toppo S, Valle G, Lanfranchi G: **Pattern recognition in gene expression profiling using DNA array: a comparative study of different statistical methods applied to cancer classification.** *Hum Mol Genet* 2003, **12:**823-836.
14. Tikhonov AN, Arsenin VY: *Solutions of ill-posed problems* Edited by: *Winston WH. Washington D.C*; 1977.
15. Wahba G: *Spline models for observational data Volume 59. CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial & Applied Mathematics*; 1990.
16. Girosi F, Jones M, Poggio T: **Regularization Theory and Neural Networks Architectures.** *Neural Computation* 1995, **7:**219-269.
17. Rifkin R, Yeo G, Poggio T: **Regularized Least Squares Classification.** In *Advances in Learning Theory: Methods, Model and Applications, NATO Science Series III: Computer and Systems Sciences Volume 190*. Edited by: *Suykens BM Horvath. Vandewalle, Amsterdam: IOS Press*; 2003:131-153.
18. Zhang P, Peng J: **SVM vs regularized least squares classification.** In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR '04): 23–26 Aug IEEE Computer Society*; 2004:176-179.
19. Luntz A, Brailovsky V: **On estimation of characters obtained in statistical procedure of recognition.** *Technicheskaya Kibernetica* 1969, **3:**.
20. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ: **Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.** *PNAS* 1999, **96:**6745-6750.
21. Ambroise C, McLachlan G: **Selection bias in gene extraction on the basis of microarray gene-expression data.** *PNAS* 2002, **99:**6562-6566.
22. Osuna E, Freund R, Girosi F: **Support Vector Machines: Training and Applications.** In *Tech Rep A.I. Memo No. 1602 Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA*; 1997.
23. Blum A, Langley P: **Selection of Relevant Features and Examples in Machine Learning.** *Artificial Intelligence* 1997, **97:**245-271.
24. Guyon I, Weston J, Barnhill S, Vapnik V: **Gene selection for cancer classification using Support Vector Machines.** *Machine Learning* 2002, **46:**389-422.
25. Good P: *Permutation tests: a practical guide to resampling methods for testing hypothesis Springer Verlag*; 1994.