Cancer recognition with bagged ensembles of Support Vector Machines

Giorgio Valentini [a] , Marco Muselli [b] and Francesca Ruffino [c]

[a] *DSI - Dipartimento di Scienze dell'Informazione*
Università di Milano, Italy.

[b] *IEIIT - Istituto di Elettronica e di Ingegneria dell'Informazione*
e delle Telecomunicazioni,
Consiglio Nazionale delle Ricerche, Genova, Italy.

[c] *ISTI - Istituto di Scienze e Tecnologie dell'Informazione*
Consiglio Nazionale delle Ricerche, Pisa, Italy.

Abstract

Expression-based classification of tumors requires stable, reliable and variance reduction methods, as DNA microarray data are characterized by low size, high dimensionality, noise and large biological variability. In order to address the variance and curse of dimensionality problems arising from this difficult task, we propose to apply bagged ensembles of Support Vector Machines (SVM) and feature selection algorithms to the recognition of malignant tissues. Presented results show that bagged ensembles of SVMs are more reliable and achieve equal or better classification accuracy with respect to single SVMs, whereas feature selection methods can further enhance classification accuracy.

Key words: Molecular classification of tumors; DNA microarray; bagging; Support Vector Machines.

1 Introduction

DNA microarray data provide a functional portrait of tumors, opening new perspectives for the classification and diagnosis of malignancies at molecular level [4].

Email addresses: valentini@dsi.unimi.it (Giorgio Valentini), marco.muselli@ieiit.cnr.it (Marco Muselli), ruffino@ieiit.cnr.it (Francesca Ruffino).

Several supervised methods have been applied to the analysis of cDNA microarrays and high density oligonucleotide chips [5,7,9]. In particular, Support Vector Machines (SVM) have been recently applied to the analysis of DNA microarray gene expression data in order to classify normal and malignant tissues and multiple tumor types [6,10]. Other approaches pointed out the importance of feature selection methods to reduce the high dimensionality of the input space [8]. In recent works, combinations of binary classifiers (one-versus-all and all-pairs) and Error Correcting Output Coding (ECOC) ensembles of MLP, as well as ensemble methods based on resampling techniques, such as bagging and boosting, have been applied to the molecular classification of tumors [5,10]. Indeed variance problems arising from small samples and biological variability of the data can be addressed through ensemble methods based on resampling techniques, while a possible way of dealing with the curse of dimensionality is offered by feature selection algorithms.
In this work we deal with these problems, combining bagged ensembles of SVMs and feature selection methods to enhance the accuracy and the reliability of malignancy predictions based on gene expression data.

2 Bagged ensembles of SVMs

We can represent the output of a single experiment with a DNA microarray as a pair $(\boldsymbol{x}, y)$, being $\boldsymbol{x} \in \mathbb{R}^d$ a vector containing the expression levels for d selected genes and $y \in \{-1, +1\}$ a binary variable determining the classification of the considered tissue. Denote with $\{\mathcal{T}_b\}_{b=1}^{B}$ a collection of B bootstrapped samples with n elements, generated by choosing at random examples in the training set $\mathcal{T} = \{(\boldsymbol{x}_j, y_j) \,:\, j = 1, \ldots, n\}$ according to a uniform probability distribution. Since the elements from $\mathcal{T}$ are drawn with replacement, every $\mathcal{T}_b$ may contain replicates. Suppose, without loss of generality, that the first n^+ pairs of $\mathcal{T}$ have $y_j = +1$, whereas the remaining $n^- = n - n^+$ possess a negative output $y_j = -1$.

Let $f_b : \mathbb{R}^d \to \mathbb{R}$ be the discriminant functions obtained by applying the soft-margin SVM learning algorithm [3] on the bootstrapped samples $\mathcal{T}_b$:

$$f_b(\boldsymbol{x}) = b + \sum_{j=1}^{n} \alpha_j y_j K(\boldsymbol{x}_j, \boldsymbol{x}) \tag{1}$$

where the scalars α_j and the bias b are obtained through the solution of a quadratic programming problem. The symmetric function $K(\cdot, \cdot)$ must be chosen among the kernels of Reproducing Kernel Hilbert Spaces [11] (e.g. a polynomial or a Gaussian).

Every f_b is associated with a decision functions $h_b : \mathbb{R}^d \to \{0, 1\}$ defined as

$h_b(\boldsymbol{x}) = \text{sign}(f_b(\boldsymbol{x}))$. In this way a set of different classifiers (*base learners*) is generated, thus exploiting the diversity of the bootstrapped samples $\mathcal{T}_b$. The generalization ability of these base learners can be improved by aggregating them through the standard majority voting formula (for two class classification problems) [2]:

$$h_{\text{st}}(\boldsymbol{x}) = \text{sign}\left(\sum_{b=1}^{B} h_b(\boldsymbol{x})\right) \tag{2}$$

Different choices of discriminant function for the bagged ensemble are possible, some of which lead to the standard decision function $h_{\text{st}}(\boldsymbol{x})$. The following three expressions allow also to evaluate the quality of the classification offered by the bagged ensemble:

$$\begin{aligned} f_{\text{avg}}(\boldsymbol{x}) &= \frac{1}{B}\sum_{b=1}^{B} f_b(\boldsymbol{x}) & f_{\text{win}}(\boldsymbol{x}) &= \frac{1}{|B^*|}\sum_{b\in B^*} f_b(\boldsymbol{x}) \\ f_{\text{max}}(\boldsymbol{x}) &= h_{\text{st}}(\boldsymbol{x}) \cdot \max_{b\in B^*} |f_b(\boldsymbol{x})| \end{aligned} \tag{3}$$

where the set $B^* = \{b : h_b(\boldsymbol{x}) = h_{\text{st}}(\boldsymbol{x})\}$ contains the indices b of the base learners that vote for the class $h_{\text{st}}(\boldsymbol{x})$. Note that $f_{\text{avg}}(\boldsymbol{x})$ is the average of the $f_b(\boldsymbol{x})$, whereas $f_{\text{win}}(\boldsymbol{x})$ and $f_{\text{max}}(\boldsymbol{x})$ are, respectively, the average of the discriminant functions of the classifiers having indices in B^* and the signed maximum of their absolute value.

The decision functions $h_{\text{win}}(\boldsymbol{x}) = \text{sign}(f_{\text{win}}(\boldsymbol{x}))$ and $h_{\text{max}}(\boldsymbol{x}) = \text{sign}(f_{\text{max}}(\boldsymbol{x}))$ are equivalent to the standard choice $h_{\text{st}}(\boldsymbol{x})$, where each base learner receives the same weight. On the contrary, with $h_{\text{avg}}(\boldsymbol{x}) = \text{sign}(f_{\text{avg}}(\boldsymbol{x}))$ the decision of each classifier in the ensemble is weighted via its prediction strength.

3 Quality assessment of classifiers

Besides the *success rate*

$$Succ = \frac{1}{2n}\sum_{j=1}^{n} |y_j + h(\boldsymbol{x}_j)| \tag{4}$$

which is an estimate of the generalization error, several alternative measures can be used to assess the quality and to evaluate the confidence of the classification performed by simple SVMs and bagged ensembles of SVMs.

By generalizing a definition introduced in [7,8], a first choice is the *extremal*

margin M_{ext}, defined as

$$M_{\text{ext}} = \frac{\theta^+ - \theta^-}{\max\limits_{1\leq j\leq n} f(\boldsymbol{x}_j) - \min\limits_{1\leq j\leq n} f(\boldsymbol{x}_j)} \tag{5}$$

being $\theta^+ = \min_{1\leq j\leq n^+} f(\boldsymbol{x}_j)$ and $\theta^- = \max_{n^++1\leq j\leq n} f(\boldsymbol{x}_j)$. It can be easily seen that the larger is the value of M_{ext}, more confident is the classifier. An alternative measure, less sensitive to outliers, is the *median margin* M_{med}:

$$M_{\text{med}} = \frac{\lambda^+ - \lambda^-}{\max\limits_{1\leq j\leq n} f(\boldsymbol{x}_j) - \min\limits_{1\leq j\leq n} f(\boldsymbol{x}_j)} \tag{6}$$

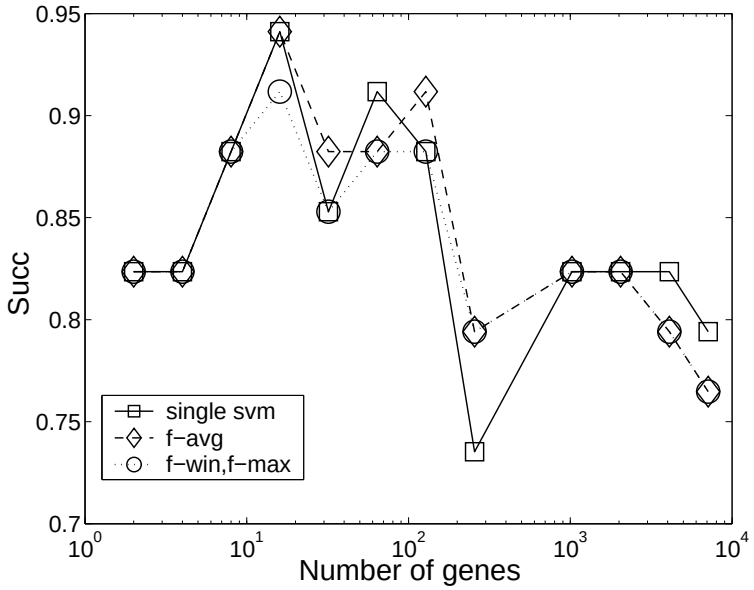
where λ^+ and λ^- are respectively the median value of $f(\boldsymbol{x})$ for the positive and negative class.

4 Numerical experiments

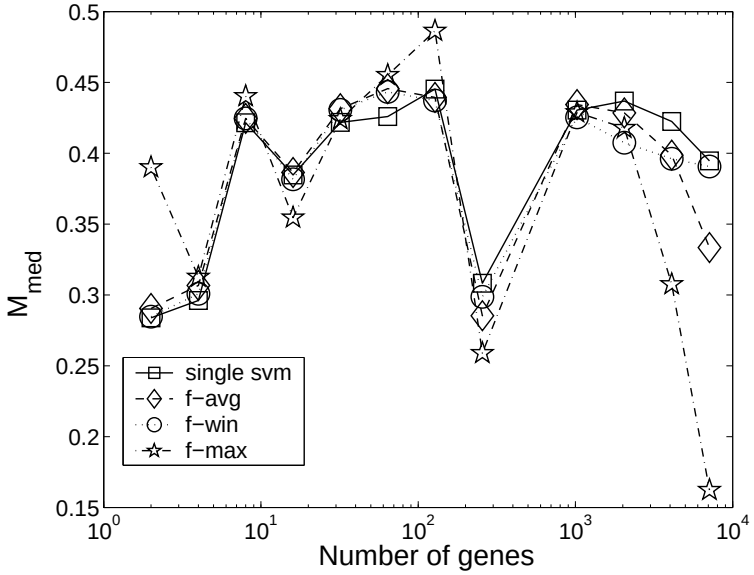
We applied SVM linear classifiers to separate normal and malignant tissues with and without feature selection. Then we compared the results obtained with single and bagged SVMs, using in all the cases the simple filter method for feature selection described in [7].

The proposed approach has been tested on the *Colon cancer* data set [1] constituted by 2000-dimensional samples including 22 normal and 40 colon cancer tissues. The whole data set has been randomly split into a training and a test set of equal size, each one with the same proportion of normal and malignant examples. We also compared the different classifiers on the *Leukemia* data set [7], which considers the problem of recognizing two variants of leukemia by analyzing the expression level of 7129 different genes. It consists of 72 examples, 47 cases of Acute Lymphoblastic Leukemia (ALL) and 25 cases of Acute Myeloid Leukemia (AML), split into a training set of 38 tissues and a test set of 34 tissues. Data preprocessing has been performed according to [1,7].

Figg. 1 and 2 compare the results obtained through the application of bagged ensembles of SVMs (for different choice of the decision function) with those achieved by single SVMs. On the *Leukemia* data set, bagging seems not to improve the success rate, even if the predictions are more reliable, especially when a small number of selected genes is used (Figg. 1a,1b). On the contrary, bagging improves the success rate scored on the *Colon* data set, both with and without feature selection, in particular if the f_{avg} discriminant function is used. (Fig. 2a).
Bagged ensembles show clearly larger median margins with respect to single



(a)

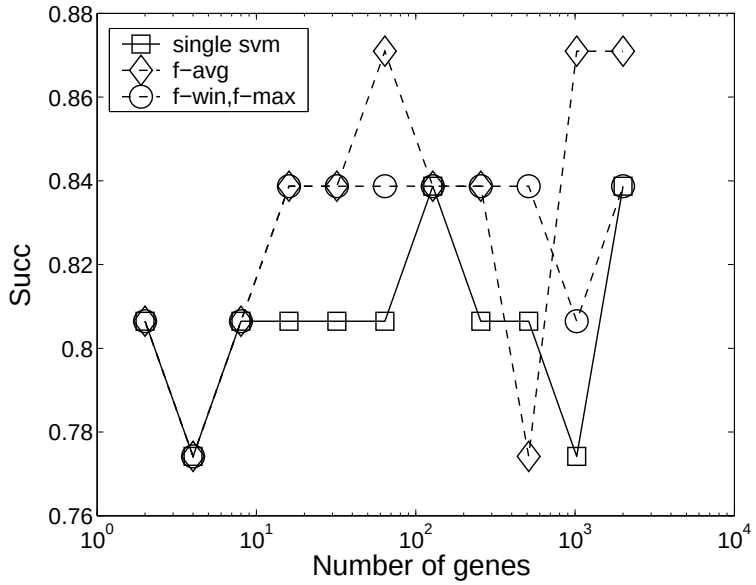


(b)

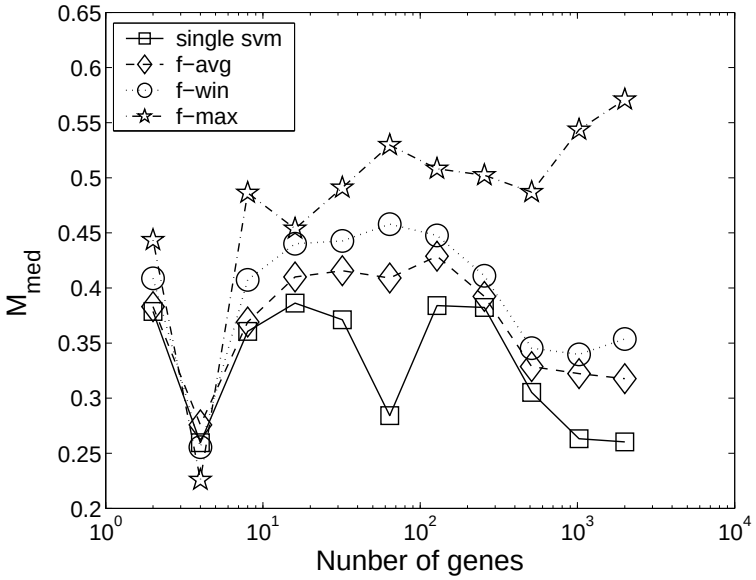
Fig. 1. *Leukemia* data set. Comparison of the results obtained with single and bagged SVMs, when varying the number of selected genes. Single SVM results are represented by continuous lines, while dotted and dashed lines represent bagged ensembles. (a) Success rate, (b) Median margin.

SVMs, confirming a better overall reliability (Figg. 1b,2b). Similar results are obtained with respect to the maximal margin (data not shown), both with the *Leukemia* and the *Colon* data set; however in *Colon* we observe an opposite behavior if the number of considered genes is relatively large.

The results show that bagged ensembles of SVMs are more reliable than single



(a)



(b)

Fig. 2. *Colon* data set. Comparison of the results obtained with single and bagged SVMs, when varying the number of selected genes. (a) Success rate, (b) Median margin.

SVMs in classifying DNA microarray data. Moreover they obtain an equivalent or a better accuracy, at least with *Colon* and *Leukemia* data sets. Anyway it is difficult to establish if a statistically significant difference between the two approaches does exist, given the small size of the available samples. Our results show also that gene selection not always can enhance the recognition rate of tumoral samples: according to [8], it plays a significant role in separating

AML from ALL, while, with the *Colon* data set, bagging, rather than feature selection, improves the accuracy and the reliability of SVMs.

References

[1] U. Alon et al. Broad patterns of gene expressions revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, 96 (1999) 6745–6750.

[2] L. Breiman. Bagging predictors. *Machine Learning*, 24(2) (1996) 123–140.

[3] C. Campbell Kernel methods: a survey of current techniques *Neurocomputing*, 48 (2002) 63–84.

[4] C.H. Chung, P.S. Bernard, and C.M. Perou. Molecular portraits and the family tree of cancer. *Nature Genetics*, 32 (2002) 533–540.

[5] S. Dudoit, J. Fridlyand, and T. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of American Statistical Association*, 97(457) (2002) 77–87.

[6] T.S. Furey et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10) (2000) 906–914.

[7] T.R. Golub et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286 (1999) 531–537.

[8] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46 (2002) 389–422,

[9] J. Khan et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7(6) (2001) 673–679.

[10] G. Valentini. Gene expression data analysis of human lymphoma using support vector machines and output coding ensembles. *Artificial Intelligence in Medicine*, 26(3) (2002) 283–306.

[11] G Wahba. Spline Models for Observational Data. *Regional Conference Series in Applied Mathematics*, vol. 59 SIAM, Philadelphia, USA, (1990).