Bio-molecular cancer prediction with random subspace ensembles of Support Vector Machines

Alberto Bertoni [a] , Raffaella Folgieri [a] and Giorgio Valentini [a]

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

Abstract

Support Vector Machines (SVMs), and other supervised learning techniques have been experimented for the bio-molecular diagnosis of malignancies, using also feature selection methods. The classification task is particularly difficult because of the high dimensionality and low cardinality of gene expression data. In this paper we investigate a different approach based on random subspace ensembles of SVMs: a set of base learners is trained and aggregated using subsets of features randomly drawn from the available DNA microarray data. Experimental results on the colon adenocarcinoma diagnosis and medulloblastoma clinical outcome prediction show the effectiveness of the proposed approach.

Key words: Molecular classification of tumors; DNA microarray; ensemble of learning machines; random subspace; Support Vector Machines.

1 Introduction

When the diagnosis of malignancies depend on multiple bio-molecular factors, traditional clinical diagnostic approaches may sometimes fail [1] High throughput bio-technologies based on large scale hybridization techniques (e.g. DNA microarray) are able to produce information to support both diagnosis and prognosis of malignancies at bio-molecular level, but the problem of extracting significant knowledge from the data becomes critical because of the peculiar

Email addresses: `bertoni@dsi.unimi.it` (Alberto Bertoni), `folgieri@dico.unimi.it` (Raffaella Folgieri), `valentini@dsi.unimi.it` (Giorgio Valentini).

Preprint submitted to Neurocomputing *28 July 2004*

characteristics of gene expression data. In fact, DNA microarray data are usually characterized by a *small number* of vectors of *high dimension* that give rise to the so called *curse of dimensionality* problem [2].

A possible approach to reduce the dimensionality consists in applying feature selection methods [3]. Considering that feature selection is a NP-hard problem [4], we experiment with an alternative approach based on random subspace ensembles [5], using linear SVMs as base learners. Indeed it has been shown that linear SVMs are effective both as base learners in random subspace ensembles [6], and as classifiers for gene expression-based diagnosis of cancer [7].

2 Random subspace ensembles for gene expression data analysis

Recently ensemble methods have been successfully applied to the bio-molecular diagnosis of tumors [8,9].

In particular random subspace ensembles seem to be well-suited to the characteristics of gene expression data, as they reduce the high dimensionality of the data by randomly selecting subsets of genes, in presence of dependencies among co-regulated genes [1]. Moreover, aggregating the resulting base classifiers trained on different subsets of gene expression levels, we may improve diversity between base learners, without a substantial loss of accuracy due to the redundancy of the available gene expression data.

At a high level, the random subspace ensemble method [5] is characterized by three steps:

(1) Given a d-dimensional data set $\mathcal{D} = \{(\boldsymbol{x}_j, t_j) | 1 \leq j \leq m\}$, $\boldsymbol{x}_j \in \mathcal{X} \subset \mathbb{R}^d$, $t_j \in \mathcal{C} = \{1, \ldots, c\}$, n new projected k-dimensional data sets $D_i = \{(P_i(\boldsymbol{x}_j), t_j) | 1 \leq j \leq m\}$ are generated $(1 \leq i \leq n)$, where P_i is a random projection $P_i : \mathbb{R}^d \rightarrow \mathbb{R}^k$. P_i is obtained by random selecting, through the uniform probability distribution, a k-subset $A = \{\alpha_1, \ldots, \alpha_k\}$ from $\{1, 2, \ldots, d\}$ and setting $P_i(x_1, \ldots, x_d) = (x_{\alpha_1}, \ldots, x_{\alpha_k})$.

(2) Each new data set D_i is given in input to a fixed learning algorithm $\mathcal{L}$ which outputs the classifiers h_i for all $i, 1 \leq 1 \leq n$.

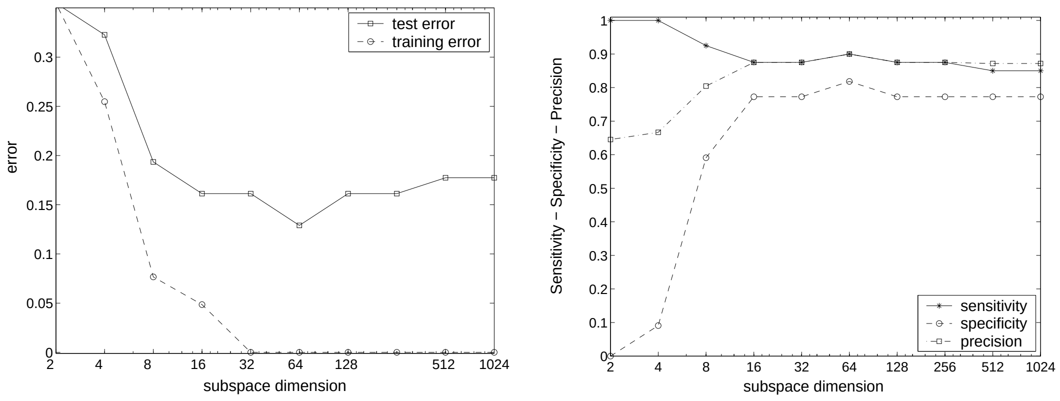
(3) The final classifier h is obtained by aggregating the base classifiers $h_1, ..., h_n$ through majority voting.

3 Experimental results

We have experimented with the method on 2 bio-medical problems, both based on gene expression profiles of a relatively small group of patients: 1) *Colon adenocarcinoma* bio-molecular diagnosis [10] 2) *Medulloblastoma* clinical outcome prediction [11]. We extended the *NEURObjects* library [12], adding new C++ classes and developing applications for random subspace ensembles.

3.1 Colon tumor prediction

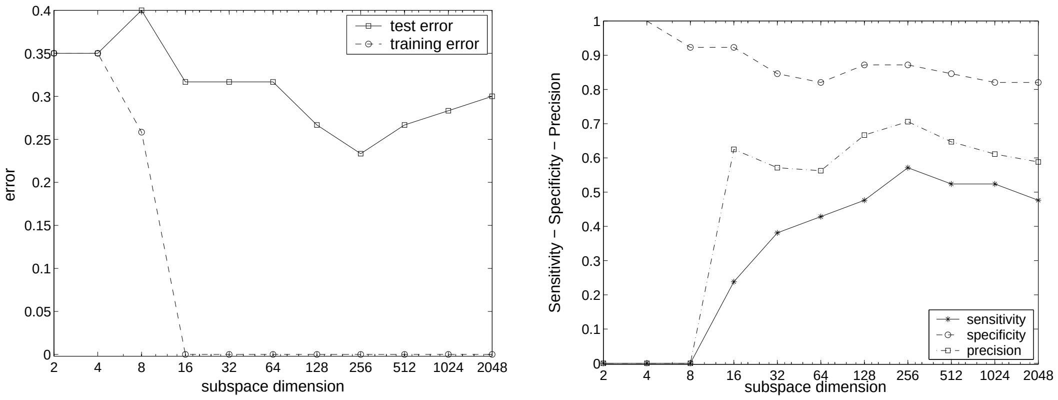
The Colon adenocarcinoma data set is composed of 2000 genes and 62 samples: 40 colon tumor samples and 22 normal colon tissue samples. We used the same preprocessing technique illustrated in [10].



(a) (b)

Fig. 1. SVM random subspace ensembles results on the colon data set (5-fold cross validation). (a) Test and training error with respect to the dimension of the subspace (b) Sensitivity, specificity and precision with respect to the dimension of the subspace.

Single linear SVMs trained using the entire set of gene expression data achieved an error of 17.74 ± 10.87 % according to a 5-fold cross validation evaluation of the generalization error. With random subspace ensembles of linear SVMs, we obtained the minimum of the test error using 64-dimensional subspaces, but also with 16 to 1024-dimensional subspaces results are equal or better than single SVMs trained on the entire feature space (Fig 1 a). The ensembles start to learn when 8 random genes are selected, and if we apply at least 16 gene-subspaces we achieve a reasonable specificity at the expense of a low decrement of the sensitivity (Fig 1 b).



(a) (b)

Fig. 2. SVM random subspace ensembles results on the medulloblastoma data set (5-fold cross validation). (a) Test and training error with respect to the dimension of the subspace (b) Sensitivity, specificity and precision with respect to the dimension of the subspace.

3.2 Medulloblastoma clinical outcome prediction

The second problem concerns with the prediction of medulloblastoma clinical outcome by gene expression profiling. The data set is composed of 60 samples, with 39 survivors and 21 treatment failures. We selected about 4000 genes from the original 7129, obtained by the same preprocessing techniques adopted in the original work [11].

Single linear SVMs trained using the entire set of gene expression data achieved an error of 28.33 ± 9.50 % according to a 5-fold cross validation estimate of the generalization error.

Random subspace ensembles outperform single SVMs trained on the entire set of the gene expression data. The minimum of the test error is registered with 256-dimensional subspaces, but in this case we need from 128 to 512-dimensional random subsets of genes to achieve better results than single SVMs (Fig 2 a). In all cases we obtained low sensitivity (slightly better for subspaces between 128 and 512 dimensions), and large specificity for a large range of randomly selected genes (Fig 2 b). Moreover our ensemble approach on medulloblastoma clinical outcome prediction achieves better accuracy and sensitivity with respect to SVMs combined with feature selection methods proposed in [11].

4 Conclusions

Random subspace ensembles outperform single SVMs on both the considered classification tasks. The null hypothesis that the random subspace ensemble has the same error rate as single SVMs is rejected at 0.05 significance level according to the 5-fold cross validated paired t-test for both the *Colon* and *Medulloblastoma* data sets.

Moreover we achieve better results with random subspace ensembles for a quite large choice of the subspace dimension (Fig. 1 and 2). Only for subspaces of very low dimension the quality of the resulting classifier is low.

The encouraging experimental results suggest to apply random subspace techniques to high-dimensional bio-molecular diagnostic problems, possibly combining the proposed ensemble approach with state-of-the-art feature selection methods.

Acknowledgments

We would like to thank the C.I.L.E.A. for providing *Avogadro* (`http://www.cilea.it`), the Linux cluster of Xeon double processors computers used in our experiments.

References

[1] A. Alizadeh, et al., Towards a novel classification of human malignancies based on gene expression, J. Pathol. 195 (2001) 41–52.

[2] R. Bellman, Adaptive Control Processes: a Guided Tour, Princeton University Press, New Jersey, 1961.

[3] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning 46 (1/3) (2002) 389–422.

[4] E. Amaldi, V. Kann, On the approximation of minimizing non zero variables or unsatisfied relations in linear systems, Theoretical Computer Science 209 (1998) 237–260.

[5] T. Ho, The random subspace method for constructing decision forests, IEEE Trans. on Pattern Analysis and Machine Intelligence 20 (8) (1998) 832–844.

[6] M. Skurichina, R. Duin, Bagging, boosting and the randon subspace method for linear classifiers, Pattern Analysis and Applications 5 (2) (2002) 121–135.

[7] T. Furey, N. Cristianini, N. Duffy, D. Bednarski, M. Schummer, D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics 16 (10) (2000) 906–914.

[8] S. Dudoit, J. Fridlyand, 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.

[9] G. Valentini, M. Muselli, F. Ruffino, Cancer recognition with bagged ensembles of Support Vector Machines, Neurocomputing 56 (2004) 461–466.

[10] U. Alon et al., Broad patterns of gene expressions revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proceeedings of the National Academy of Sciences 96 (1999) 6745–6750.

[11] S. Pomeroy et al., Gene Expression-Based Classification and Outcome Prediction of Central Nervous System Embryonal Tumors, Nature 415 (2002) 436–442.

[12] G. Valentini, F. Masulli, NEURObjects: an object-oriented library for neural network development, Neurocomputing 48 (1–4) (2002) 623–646.