FEATURE SELECTION COMBINED WITH RANDOM SUBSPACE ENSEMBLE FOR GENE EXPRESSION BASED DIAGNOSIS OF MALIGNANCIES

Alberto Bertoni, [1] Raffaella Folgieri, [1] Giorgio Valentini, [1]

[1] *DSI, Dipartimento di Scienze dell' Informazione,*
Università degli Studi di Milano,
Via Comelico 39, 20135 Milano, Italia.

bertoni@dsi.unimi.it

folgieri@dico.unimi.it

valentini@dsi.unimi.it

Abstract The bio-molecular diagnosis of malignancies represents a difficult learning task, because of the high dimensionality and low cardinality of the data. Many supervised learning techniques, among them support vector machines, have been experimented, using also feature selection methods to reduce the dimensionality of the data. In alternative to feature selection methods, we proposed to apply random subspace ensembles, reducing the dimensionality of the data by randomly sampling subsets of features and improving accuracy by aggregating the resulting base classifiers. In this paper we experiment the combination of random subspace with feature selection methods, showing preliminary experimental results that seem to confirm the effectiveness of the proposed approach.

Keywords: Molecular diagnosis, ensemble methods, Support Vector Machine, Random Subspace, DNA microarray

1. Introduction

High throughput bio-technologies based on large scale hybridization techniques (e.g. DNA microarray) can provide information for supporting both diagnosis and prognosis of malignancies at bio-molecular level [Alizadeh, A. et al., 2001]. Several supervised methods have been applied to the analysis of cDNA microarrays and high density oligonucleotide chips (see e.g. [Dudoit et al., 2002]). The high dimensional-

ity and low cardinality of gene expression data, together with the high sensitivity required for diagnostic problems, makes the classification of malignant and normal samples very challenging from a machine learning point of view.

An effective approach to this problem is represented by feature selection methods [Guyon et al., 2002], that can be useful both to select the genes more related to malignancies and to enhance the discrimination power between normal and malignant tissues. Recently we proposed an alternative approach [Bertoni et al., 2004] based on random subspace ensembles [Ho, 1998], that is sets of learning machines trained on randomly chosen subspaces of the original input space.

In this paper we propose to integrate the two approaches in order to enhance the accuracy and the reliability of the diagnostic system: at a first stage a subset of genes is selected through a feature selection method, successively subsets of genes randomly drawn from the previously selected genes are used to train an ensemble of learning machines. The ensemble output can be obtained through majority voting or any other aggregation technique. We call this method *Random Subspace on Selected Features* (*RS-SF*).

The proposed combined approach is described in the next section. Some preliminary experimental results are shown in Sect. 3, while in the last section we report conclusions and on-going developments of the present work.

2. Feature selection methods and random subspace ensembles for gene expression data analysis

The major problem in gene expression analysis is the high dimensionality and low cardinality of the data, from which the curse of dimensionality problem arises.

An approach to this problem consists in reducing the dimensionality through feature (gene) selection methods [Golub et al., 1999; Guyon et al., 2002]. Many methods can be applied, ranging from filter methods, wrapper methods, information theory based techniques and "embedded" methods (see e.g. [Guyon and Elisseeff, 2003] for a recent review).

On the other hand we recently experimented a different approach [Bertoni et al., 2004] based on random subspace ensemble methods [Ho, 1998]. For a fixed k, k-subsets of features are selected. according to the uniform distribution. Then the data of the original training set are projected to the selected n-dimensional subspaces and the resulting data sets are used to train an ensemble of learning machines [Ho, 1998].

RS-SF Algorithm

`Input`:
- A data set $\mathcal{D} = \{(\mathbf{x}_j, t_j)|1 \leq j \leq m\}$, $\mathbf{x}_j \in \mathcal{X} \subset \mathbb{R}^d$, $t_j \in \mathcal{C} = \{1, \ldots, k\}$
- a learning algorithm $\mathcal{L}$
- a feature selection algorithm $\mathcal{F}$
- a number of selected features $n < d$
- a dimension $k < n$ of the random subspace
- number of the base learners I

`Output`:
- Final hypothesis $h_{ran} : \mathcal{X} \to \mathcal{C}$ computed by the ensemble.

`begin`
$\hat{\mathcal{D}} = \mathcal{F}(\mathcal{D}, \text{n})$
for $i = 1$ to I
begin
$D_i =$ `Subspace_projection`$(\hat{\mathcal{D}}, \text{k})$
$h_i = \mathcal{L}(D_i)$
end
$h_{ran}(\mathbf{x}) = \arg\max_{t \in \mathcal{C}} card(\{i|h_i(\mathbf{x}) = t\})$
`end`.

Figure 1. Random Subspace on Selected Features (*RS-SF*) ensemble method.

In this work we experiment a combination of the two approaches. The role of the gene selection stage consists in eliminating noisy or uninformative genes. Then we can apply random subspace ensembles only with the remaining more discriminant and informative genes, enhancing the accuracy of the resulting base learners though aggregation techniques, while diversity between base learners is maintained by the random choice of the input subspaces.

Fig. 1 summarizes the proposed method. $\mathcal{F}$ denotes a feature selection algorithm, that selects the n most significant features from the original d-dimensional input space. `Subspace_projection` is a randomized procedure that selects, according to the uniform distribution, a k-subset $A = \{\alpha_1, \ldots, \alpha_k\}$ from $\{1, 2, \ldots, n\}$, so defining a projection $P_A : \mathbb{R}^n \to \mathbb{R}^k$, where $P_A(x_1, \ldots, x_n) = (x_{\alpha_1}, \ldots, x_{\alpha_k})$; then it returns as output the new k-dimensional data set $\{(P_A(\mathbf{x}_j), t_j)|1 \leq j \leq m\}$, where $\hat{\mathcal{D}} = \{(\mathbf{x}_j), t_j)|1 \leq j \leq m\}$ is the set of the n-dimensional features selected from the original d-dimensional input space. Every new data set D_i obtained through the iteration of the procedure `Subspace_projection` is given as input to a learning algorithm $\mathcal{L}$ which outputs a classifier h_i. Note that, with

abuse of notation, with $h_i(\mathbf{x})$ we ambiguously denote the extension of h_i to the entire $\mathbb{R}^d$ space. All the obtained classifiers are finally aggregated through majority voting.

3. Experiments with the colon adenocarcinoma gene expression data

To evaluate the feasibility of the *RS-SF* ensemble method for the analysis of gene expression data, we considered the colon adenocarcinoma bio-molecular diagnosis problem. The *colon* data set is composed of 62 samples: 40 colon tumor and 22 normal colon tissue samples, with 2000 gene expression data for each sample [Alon, U. et al., 1999].

Main goal of the experiment is the performance comparison of SVMs trained with subsets of genes chosen through a simple but effective feature selection method (Golub's method) [Golub et al., 1999] and *RS-SF* ensembles.

3.1 Experimental setup

Regarding preprocessing of data, we used the same techniques illustrated in [Alon, U. et al., 1999]. Groups of genes have been selected ranking the gene's scores obtained through the Golub's statistics. The selection of the genes has been performed using only training data in order to avoid the selection bias [Ambroise and McLachlan, 2002].

Table 1. Summary of the best results achieved with single SVMs trained on subsets of genes selected through Golub's method (Single *FS-SVM*), *RS-SF* ensembles of SVMs, standard random subspace ensembles (*RS* ensemble), single SVMs without feature selection, and the average error of the base SVMs that compose the ensemble.

	Test Err.	**St.dev**	**Train Err.**	**St.dev**	**Sens.**	**Spec.**
RS-SF ensemble	0.0968	0.0697	0.0727	0.0183	0.9250	0.8636
RS ensemble	0.1290	0.0950	0.0000	0.0000	0.9000	0.8182
Single *FS-SVM*	0.1129	0.0950	0.0768	0.0231	0.9250	0.8182
Single *SVM*	0.1774	0.1087	0.0000	0.0000	0.8500	0.7727
Single base *SVM*	0.1776	0.1019	0.0000	0.0000	——	——

We considered different random subspaces of dimensionality from 2 to 2^{n-1}, randomly drawn from each 2^n-dimensional gene space selected from the input space through the Golub's method, while varying n between 5 and 10. According to Skurichina e Duin [Skurichina and Duin, 2002] we applied linear SVMs as base learners. Indeed they showed that random subspace ensembles are effective with linear base learners characterized by a decreasing learning curve (error) with respect to the

cardinality n, especially when the dimensionality is much larger than the cardinality. For each ensemble we trained 200 linear SVMs, considering values of the regularization parameter C between 0.01 and 1000.

We computed for both single SVMs and *RS-SF* ensembles the test error and training error, sensibility, specificity and precision through 5-fold cross validation techniques. Regarding software, we developed new C++ classes and applications for random subspace ensembles extending the *NEURObjects* library [Valentini and Masulli, 2002]. For all the experiments, we used the C.I.L.E.A. Avogadro cluster of Xeon double processor workstations [Arlandini, 2004].

3.2 Results

The results show that the best *RS-SF* ensemble outperforms single SVMs trained with a subset of selected genes (Single *FS-SVM*). In fact we obtained respectively a 0.0968 test error in the first case, and 0.1129 for *FS-SVM* (Tab. 1). The test error of *RS-SF* ensemble is consistently equal or lower than single *FS-SVM*, independently of the number of the selected genes, as shown in Fig. 2. In particular the minimum of the test error with 128 selected genes is obtained with 64-dimensional random subspace, while with 512 selected genes with 16-dimensional subspaces. In both considered methods, the sensitivity has the same value from 32 to 128 selected genes, then it decreases for single *FS-SVM*, while becomes constant for *RS-SF* ensembles (Fig. 3). Also the specificity is better for random subspace ensemble combined with feature selection: a maximum is achieved with 128 selected genes, and for number of selected genes larger than 64 *RS-SF* ensembles show better results than single *FS-SVM* (Fig. 3).

The difference between the best *RS-SF* ensemble and single *FS-SVM* is not statistically significant, according to the 5-fold cross validated t-test [Dietterich, 1998] (Tab. 1). On the contrary it becomes significant with standard random subspace ensemble and single SVMs trained without feature selection. Anyway, considering the accuracy in *RS-SF* ensemble and single *FS-SVM* with respect to the number of the selected genes, the difference is significant at 0.05 level in most cases (Fig. 2).

4. Conclusions and developments

The results show the applicability of the combined approach of the random subspace ensemble with feature selection methods, to the analysis of gene expression data. Anyway we need to perform more experiments with other data sets, to confirm, as may be expected, the presented results. The proposed approach doesn't require a specific fea-

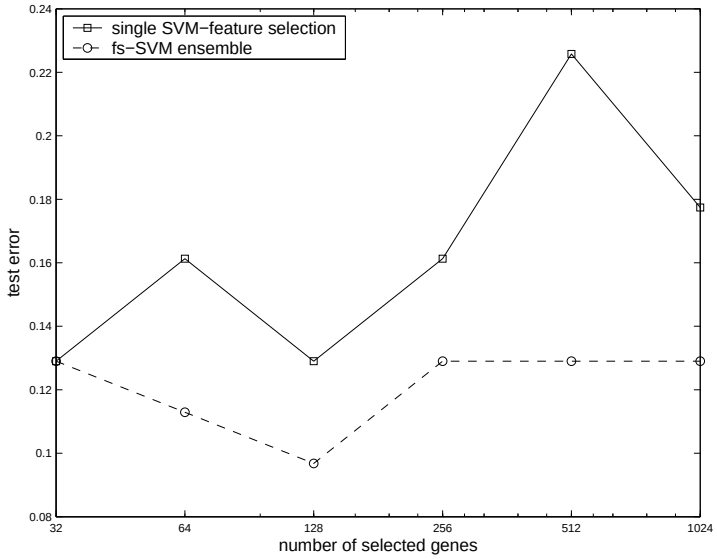

Figure 2. Comparison of the test error with respect to the number of the selected features between *RS-SF* ensembles of SVMs and *FS-SVMs*.

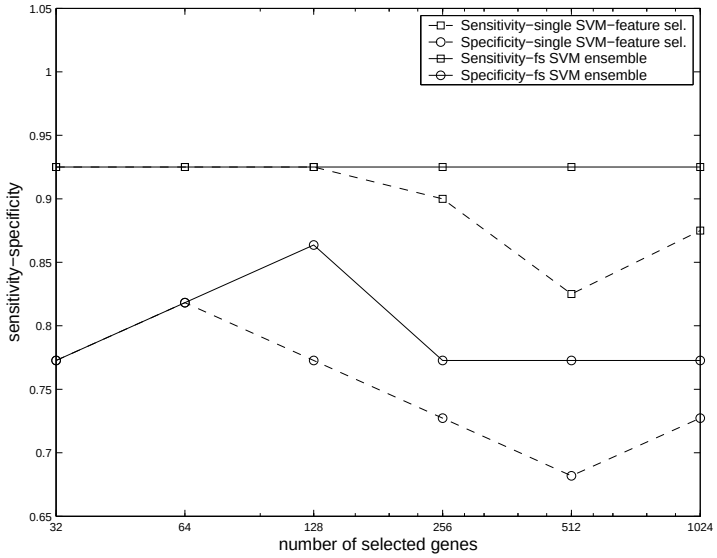

Figure 3. Comparison of sensitivity and specificity with respect to the number of the selected features between *RS-SF* ensembles of SVMs (continuous lines) and *FS-SVMs* (dashed lines).

ture selection method. Regarding this item, we plan to experiment with other feature selection algorithms.

Acknowledgments

We would like to thank the C.I.L.E.A. for providing *Avogadro* [Arlandini, 2004], the computer Linux cluster used in our experiments.

References

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

Alon, U. et al. (1999). Broad patterns of gene expressions revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS*, 96:6745–6750.

Ambroise, C. and McLachlan, G. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *PNAS*, 99(10):6562–6566.

Arlandini, C. (2004). Avogadro: il CILEA oltre il muro del teraflop. *Bollettino del CILEA*, 91.

Bertoni, A., Folgieri, R., and Valentini, G. (2004). Random subspace ensembles for the bio-molecular diagnosis of tumors.

Dietterich, T.G. (1998). Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation*, (7):1895–1924.

Dudoit, S., Fridlyand, J., and Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *JASA*, 97(457):77–87.

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

Guyon, I. and Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157–1182.

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

Ho, T.K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844.

Skurichina, M. and Duin, R.P.W. (2002). Bagging, boosting and the randon subspace method for linear classifiers. *Pattern Analysis and Applications*, 5(2):121–135.

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