BICA AND RANDOM SUBSPACE ENSEMBLES FOR DNA MICROARRAY-BASED DIAGNOSIS

B. APOLLONI AND G. VALENTINI

Dipartimento di Scienze dell'Informazione,
Università degli Studi di Milano,
Via Comelico 39/41, 20135 Milano, Italy
{*apolloni, valentini*}*@dsi.unimi.it*

A. BREGA

Dipartimento di Matematica "F. Enriquez",
Università degli Studi di Milano
Via Saldini 50, 20133 Milano, Italy
andrea.brega@unimi.it

We compare two ensemble methods to classify DNA microarray data. The methods use different strategies to face the course of dimensionality plaguing these data. One of them projects data along random coordinates, the other compresses them into independent boolean variables. Both result in random feature extraction procedures, feeding SVMs as base learners for a majority voting ensemble classifier. The classification capabilities are comparable, degrading on instances that are acknowledged anomalous in the literature.

1. Introduction

The traditional taxonomy of malignancies, based on their morphological, histopathological, and clinical characteristics, may be sometimes ineffective for a correct diagnosis and prognosis of tumors [1]. Indeed a more refined diagnosis may be achieved exploiting the genome-wide bio-molecular characteristics of tumors, using high throughput bio-technologies based on large scale hybridization techniques (e.g. DNA microarray) [5].

One of the main drawbacks that characterizes DNA microarray data is represented by their very high dimensionality and low cardinality. Indeed is well known that in these cases the *curse of dimensionality* problem arises. Hence several works pointed out the importance of feature selection methods to reduce the dimensionality of the input space [7]. An alternative approach is represented by data compression techniques that can reduce the

dimensionality of the data, while approximately preserving their information content. As for their processing, several authors recently proposed to apply ensemble methods for improving the performance of state-of-the-art classification algorithms in the context of gene expression data analysis [4].

In this paper we compare two ensemble methods based on data-compression techniques for DNA-microarray-based diagnosis. The first one exploits random projections to lower dimensional subspaces [8], while the second performs data compression through a Boolean Independent Component Analysis (BICA) algorithm [13]. While the first method has just been applied to gene expression data analysis [3], BICA has never been previously applied to DNA microarray data analysis.

In the next two sections we introduce the methods, and in Sect. 4 we experimentally analyze the effectiveness of the two approaches, applying them to DNA microarray-bases diagnosis of tumors.

2. RSE: Random Subspace Ensemble

The reduction of the dimensionality in the context of supervised analysis of data is usually pursued through feature selection methods. Many methods can be applied, ranging from filter methods, wrapper methods, information theory based techniques and "embedded" methods (see e.g. [6] for a recent review).

We recently experimented a different approach [3] based on random subspace ensemble methods [8]. For a fixed n, n features (genes) are randomly selected, according to the uniform distribution. Then the data of the original d-dimensional training set is projected to the selected n-dimensional subspace. The resulting data set is used to train a suitable base learner and this process is repeated ν times giving raise to an ensemble of ν learning machines trained on different randomly selected subsets of features. The resulting set of classifiers are then combined by using majority voting.

This method avoids some computational difficulty of feature selection (feature selection is an NP-hard problem), and a parallel implementation can be provided in a natural way. Anyway feature selection methods can explicitly select sets of relevant features, while this information cannot be directly obtained through RS ensembles. On the other hand, with different random projections of the data we can improve diversity between base learners [9], while the overall accuracy of the ensemble can be enhanced through aggregation techniques. As a consequence the performance of a given classification algorithm may be enhanced. A high-level pseudo-code

of the method is summarized in Fig. 1. In particular, `Subspace_projection`

Random Subspace Ensemble 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 $\mathcal{L}$
- subspace dimension $n < d$
- number of the base learners m

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

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

Figure 1. High-level pseudo-code of the RSE method

procedure selects a n-subset $A = \{\alpha_1, \ldots, \alpha_n\}$ from $\{1, 2, \ldots, d\}$, and returns as output the new data set $D_i = \{(P_A(\mathbf{x}_j), t_j) | 1 \leq j \leq m\}$, where $P_A(x_1, \ldots, x_d) = (x_{\alpha_1}, \ldots, x_{\alpha_n})$.

The new data set D_i is then given as input to a learning algorithm $\mathcal{L}$ which outputs a classifier h_i. All the classifiers obtained are finally aggregated through majority voting, where $card()$ measures the cardinality of a set.

3. BICA network

A suitable way of taking decisions based on data is to split the decision process in two steps. The first is devoted to preprocessing data in a feasible way such that they can be interpreted in the second one. As for the former, it mirrors real vectors into boolean ones, that should reflect relevant features of the original data patterns. Stressing the fact that independence is a property of the representation of the data that we use, we search for this property precisely on a concise Boolean representation of them suitable for their correctly partition into positive and negative inputs of our decision rule. Accordingly, we call the mirroring method Boolean Independent Component Analysis, BICA for short.

3.1. *The architecture*

We split the mirroring of the original data into the target Boolean vector in two parts: a true mirroring of the patterns and a projection of a compressed representation of them (obtained as an aside result of the first part) into the space of Boolean assignments. The whole process is done by a neural network with an architecture shown in Fig. 2 sharing the same input and hidden layer with the two output segments A and B computing the Boolean assignments and a copy of the input, respectively.

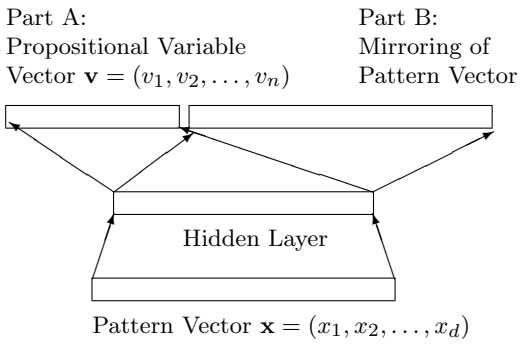

Figure 2. Layout of the neural network mapping features to symbols.

3.2. *The learning algorithm*

We train this network with a backpropagation algorithm [10] as follows.
Error backpropagation in part B. As customary with this functionality [11], we structured our network as a three-layer network with the same number of units in both input and output layers and a smaller number of units in the hidden layer. Therefore the hidden layer constitutes a bottleneck which collects in the state of its nodes a compressed representation of the input. This part of the network is trained according to a quadratic error function and usual formulas[12].
Error backpropagation in part A. Things are different for the units of part A of the output. In this case we require that the network minimizes the following error:

$$E_s = \ln \left(\prod_{k=1}^{n} z_{s,k}^{-z_{s,k}} (1 - z_{s,k})^{-(1-z_{s,k})} \right) \tag{1}$$

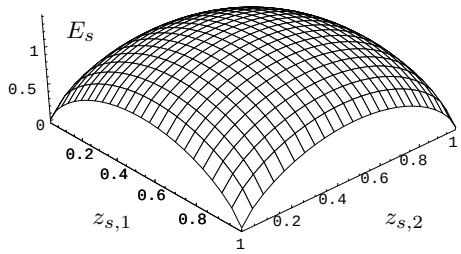

Figure 3. Graph of the function E_s with $n = 2$.

where $z_{s,j}$ is the output of the unit j upon presentation of s-th pattern. This function, which we call the *edge pulling function*, has the shape of an entropy measure that finds its minima in the vertices of the neural network output space (see Fig. 3).

The error which is backpropagated from the units of part A is:

$$\delta_{s,k} = f'_{\text{act}}(\text{net}_{s,k})\alpha_{s,k} \tag{2}$$

where $\text{net}_{s,j}$ is a weighted sum of the inputs to j-th unit on s-th pattern, f_{act} is the sigmoid function, and

$$\alpha_{s,k} = -\frac{\partial E_s}{\partial z_{s,k}} = \ln\left(\frac{z_{s,k}}{1 - z_{s,k}}\right) \tag{3}$$

In addition,we insert a syntactic feedback into eq. 3 through an extra term which has the form of a 'directed noise' $\theta_{s,k}$ added to the initial value of α when we are not satisfied with the 'correctness' of the result. Namely, when the Hamming distance between vectors corresponding to patterns belonging to different classes falls below a given threshold, we assume patterns with the minority label incorrect. Then, denoting $\tau_{s,k}$ the specific punishment to the neuron k for an incorrect pattern s, $\theta_{s,k}$ reads:

$$\theta_{s,k} = (1 - 2\Gamma(z_{s,k}))\,\tau_{s,k} \tag{4}$$

where Γ is a threshold function. The first term in the brackets specifies the sign of $\theta_{s,k}$ so that the contribution to the network parameters is in the opposite direction from the one the unit is moving in. Finally, using a tuning parameter π_A to balance parts B and A, $\alpha_{s,k}$ reads:

$$\alpha_{s,k} = \pi_A\left(\theta_{s,k} + \ln\left(\frac{z_{s,k}}{1 - z_{s,k}}\right)\right) \tag{5}$$

The joint goal of minimizing E_s and maintaining patterns well separated in two categories brings the Boolean assignments to figure as samples of independent random variables, thus we may say that these variables are *expectedly* independent. More precisely, the following lemma has been proved in [13]:

Lemma 3.1. *With reference to the neural network and training algorithm described above, if the neural network outputs are correct and all close to the vertices of the Boolean hypercube then their values stretched to the vertices constitute assignments to expectedly independent Boolean variables.*

We repeat ν times also this process getting different maps, as a consequence of the random initialization of the network parameters, and different base learners trained on the encoded training sets. Finally, we compute for each sample of the training set the frequency with which base learners answer 1, and we gather frequencies corresponding to either positive or negative samples. In the lose assumption that frequencies in each group follow a Gaussian distribution we locate a threshold at the cross of their p.d.f.s [14], i.e.

$$t = \frac{\hat{\mu}_- \hat{\sigma}_+ + \hat{\mu}_+ \hat{\sigma}_-}{\hat{\sigma}_- + \hat{\sigma}_+}$$

where $\hat{\mu}_-$ and $\hat{\sigma}_-$ are the sample estimate of parameter μ_- and σ_- of the negative distribution; idem for the positive distribution. With this threshold we classify test set records giving label 1 to those whose 1 frequency according to trained base learners overcome the threshold.

4. Colon Tumor Classification

4.1. *Experimental setup*

In order to compare the two approaches, we applied the two ensemble methods to the classification of DNA microarray data relative to colon tumor samples [2]. The Colon adenocarcinoma data set is composed of 2000 genes and 62 samples: 40 colon tumor samples and 22 normal colon tissue samples. We evaluated the generalization performances of the two ensembles using multiple hold-out techniques: we randomly split the data in two equally-sized training and test sets (i.e. 31 samples in the training and 31 samples in the test set), repeating this process 50 times. Then the average error on the test set has been computed. In both ensembles we used 60 Support Vector Machines (SVMs) as base learners. With RS ensembles

we applied different projections into random subspace with dimension from 16 to 1024, and used linear SVMs, tuning their regularization parameter. With BICA network we mapped from $\mathbb{R}^{2000}$ to $\{0,1\}^{20}$ space, and we used a second order kernel SVMs, as a result of a model selection procedure.

4.2. *Results and discussion*

Comparing the results obtained with the ensemble methods with those obtained with single SVMs, we can register a significant enhancement achieved with the ensemble approach w.r.t. the single SVMs (Tab. 1). On the other

Table 1. Classification accuracy in single SVMs, BICA and RSE ensembles.

	single SVM	RSE	BICAe
test accuracy	0.67	0.828	0.792
σ	0.07	0.05	0.09
train accuracy	0.80	1.000	0.98
σ	0.07	0.00	0.02

hand, there is no a substantial difference between the performances of the two ensemble approaches, with only a slight improvement obtained with RSE ensembles. In order to understand if the errors of BICA and RSE ensembles are approximately distributed on the same examples, we also analyzed the frequencies of their errors in function of the pattern examined across the 50 test sets used in the multiple hold-out experiments. Fig. 4 reports the compared test error frequencies by patterns. Interestingly enough, the two ensemble methods show their largest errors on the same examples (apart a few discrepancies). The largest errors are concentrated on samples 45, 49, 51, 55 and 56 for both the ensemble methods. As explained in [2], most normal samples are enriched in muscle cells, while tumor samples are enriched in epithelial cells. The above samples consistently misclassified by both ensemble methods present an "inverted" tissue composition: normal samples are rich in epithelial cells, tumor samples are rich in muscle cells. This fact shows that the separation between normal and tumoral samples is also made on the basis of tissue composition, as observed in [7].

The best results with RSE have been obtained through random projections into 64-dimensional subspaces. BICA requires only 20 bits. As a matter of fact both encodings do not represent a real strong compression of DNA data, since we need 60 different maps to obtain a satisfactory clas-

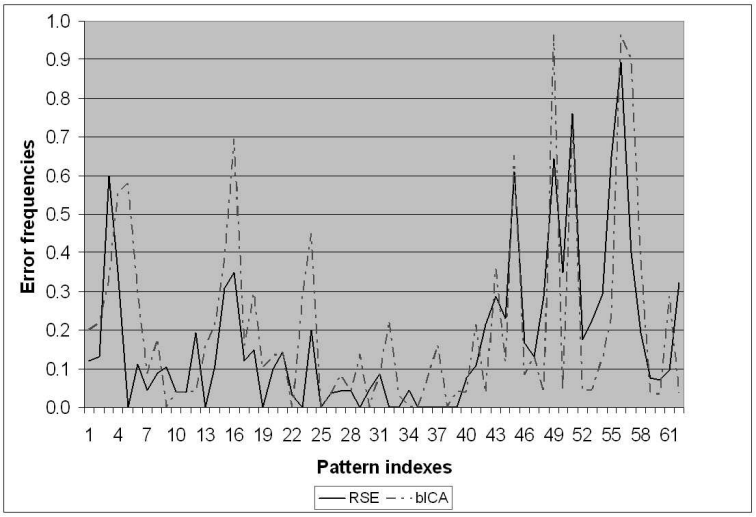

Figure 4. Frequencies of error of BICA and RSE ensemble methods in function of the pattern examined.

sification. We note however that the 63% of the database is well classified using a single SVM and 93% using only 3 SVMs. Moreover, only 14 variables are used by the mentioned single SVM involving in own turn only 151 features uniformly distributed within the topology of the bench of the 2000 features supplied by the micro-array.

These results suggest that BICA technique could be in perspective applied to discover genes relevant for tumor discrimination that may be validated by the RSE ensembles.

Acknowledgments

We would like to thank the anonymous reviewers for their comments and suggestions.

References

1. A. Alizadeh et al. Towards a novel classification of human malignancies based on gene expression. *J. Pathol.*, 195:41–52, 2001.
2. U. Alon, et al. Broad patterns of gene expressions revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS*, 96:6745–6750, 1999.
3. A. Bertoni, R. Folgieri, and G. Valentini. Bio-molecular cancer prediction

with random subspace ensembles of support vector machines. *Neurocomputing*, 63C:535–539, 2005.
4. S. Dudoit, J. Fridlyand, and T. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *JASA*, 97(457):77–87, 2002.
5. M. Eisen and P. Brown. DNA arrays for analysis of gene expression. *Methods Enzymol.*, 303:179–205, 1999.
6. I. Guyon and A. Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.
7. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1/3):389–422, 2002.
8. T.K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on PAMI*, 20(8):832–844, 1998.
9. L.I. Kuncheva and C.J. Whitaker Measures of diversity in classifier ensembles. *Machine Learning*, 51:181–207, 2003.
10. D.E. Rumelhart, G.E. Hinton and R.J. Williams. Learning Internal Representations by Error Propagation. *MIT Press, Cambridge*, 318–362, 1987.
11. J. Pollack. Recursive distributed representation. *Artificial Intelligence*, 46:77–105, 1990.
12. C.M. Bishop. Neural networks for pattern recognition. *Clarendon Press, Oxford*, 1995.
13. B. Apolloni, A. Esposito, D. Malchiodi, C. Orovas, G. Palmas, J.J Taylor. A general framework for learning rules from data. *IEEE Trans. on Neural Networks*, 11:6, 2004.
14. R.O. Duda,P.E. Hart. Pattern classification and scene analysis. *John Wiley & Sons, New York*, 1973