ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

A new method of DNA probes selection and its use with multi-objective neural networks for predicting the outcome of breast cancer preoperative chemotherapy

R. Natowicz[1], A. P. Braga[2], R. Incitti[3], E. G. Horta[2]
R. Rouzier[4], T. S. Rodrigues[5] and M. A. Costa[2]

1- Université Paris-Est, ESIEE-Paris, Département d'informatique, France

2- Universidade Federal de Minas Gerais, Depto. Engenharia Eletrônica, Brazil

3- Institut Mondor de Médecine Moléculaire, Plate-forme génomique, France

4- Hôpital Tenon, Service de gynécologie, France

5- Universidade Federal de Lavras, Depto. Ciência da Computação, Brazil

Abstract. DNA microarrays technology has emerged as a major tool to explore cancer biology and solve clinical issues. The response to chemotherapy represents such an issue because its prediction would make it possible to give the patients the most appropriate chemotherapy regimen. We propose a new method of probes selection, and we study the performances of predictors designed with multi-objective neural network (MOBJ-NN) taking as input the expression levels of the selected probes. The novelty of this paper is to link the method of probes selection and the MOBJ-NN model for designing multi-gene predictors.

1 Introduction

The development of post-genomic high-throughput measurement technologies and the associated computational analysis tools give the opportunity to identify for each tumor, a profile based on level of mRNA expression. In breast cancer, neoadjuvant chemotherapy (treatment given prior to surgery) makes it possible to check, in vivo, breast tumor chemosensitivity. A pathologic complete response (PCR) at surgery is correlated with an excellent outcome while residual disease (NoPCR) is associated with a poor outcome. An accurate prediction of tumor sensitivity to preoperative chemotherapy is an important issue because patients with predicted residual disease may avoid the prescription of an inefficient treatment and may be allocated to other treatments. The design of multigene predictors of the patients' class, PCR or NoPCR, is a supervised learning problem. The methods that are the most commonly used for selecting a subset of DNA probes are based on the identification of probes that depart the most from a random distribution of expression levels. The DNA probes are ranked and the genes are selected according their p-values of a t-test. In these methods, the forthcoming classifier models are not involved in the selection process (such mathod are often said to be part of a 'filtering approach'). In some other studies the classifier model is involved in the process of DNA probes se-

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

lection ('wrapper approach'). In the present article we present a new method of probe selection in a filtering approach, and the performance of predictors taking

2 Experimental data

The clinical trial from which the data were collected has been conducted at the Nellie B. Connally Breast Center, University of Texas M.D. Anderson Cancer Center. It is described in details in [1]. One hundred thirty-three patients with stage I-III breast cancer were included in the trial conducted at the MD Anderson Cancer Center in Houston (USA) for 82 patient cases and at the Institut Gustave Roussy in Villejuif (France) for 51 cases. The pretreatment gene expression profiling was performed with oligonucleotide microarrays (Affymetrix U133A, made out of 22283 DNA probes) on fine-needle aspiration specimens. The *training set* was composed of the former 82 patient cases and the independant *validation set* was the latter set of 51cases. At the completion of neoadjuvant chemotherapy, all the patients had surgical resection of the tumor bed, with negative margins. The pathologic complete response, PCR, was defined as no histopathologic evidence of any residual invasive cancer cells in the breast, and the non pathologic complete response, NoPCR, was defined as any residual cancer cells after histopathologic study. All the data of the clinical trial are available online at URL http://www.bioinformatics.mdanderson.org/pubdata.html.

3 Method of probes selection

We assigned two sets of expression levels to any probe s, the sets $E_p(s)$ and $E_n(s)$, computed from the training data as follows [2]. Let $m_p(s)$ and $sd_p(s)$ be the mean and standard deviation of the expression levels of probe s for the PCR training cases, and let $m_n(s)$ and $sd_n(s)$ be that of the NoPCR training cases. The set of expression levels of the PCR training cases was defined as the set difference $E_p(s)$,

$$E_p(s) = [m_p(s) - sd_p(s), m_p(s) + sd_p(s)] \setminus [m_n(s) - sd_n(s), m_n(s) + sd_n(s)]$$

and conversely for the NoPCR training cases,

$$E_n(s) = [m_n(s) - sd_n(s), m_n(s) + sd_n(s)] \setminus [m_p(s) - sd_p(s), m_p(s) + sd_p(s)].$$

Discrete probes' predictions. For any patient case, the individual prediction of a probe was a discrete value in set $\{pcr, nopcr, unspecified\}$: *pcr* if the expression level of patient p lied within the interval $E_p(s)$ and *nopcr* if it lied within $E_n(s)$. Otherwise, the individual prediction value was $unspecified$.

Probes' valuation function. Let $p(s)$ be the number of PCR training cases correctly predicted *pcr* by probe s, and let $n(s)$ be the number of the NoPCR training cases correctly predicted *nopcr* by the probe. The valuation function of the probes was defined so as to favor probes which correctly predicted high numbers of training cases and moreover, whose sets of correctly predicted training

cases were 'good' samplings of the training set. To this end, we have considered the ratios $p(s)/P$ and $n(s)/N$ of correctly predicted training cases. The valuation function $v(s)$, $v(s) \in [0, 1]$, was defined as:

$$v(s) = 0.5 \times \left(\frac{p(s)}{P} + \frac{n(s)}{N} \right).$$

4 Classifiers used as predictors

In [1], K. Hess & al. selected the probes according to the p-value of a t-test. With this probe selection method, they found that the best predictor for the dataset at hand was made out of 30 probes, whose expression levels were weighted after the result of a diagonal linear discriminant analysis (predictor 'DLDA-30'). In the present article, we have considered the 30 probes of highest values $v(s)$ (see [2] for an extensive analysis of the performances in function of the number of selected probes), and the performances of several classifiers: majority decision, support vector machine (SVM) [3], multi-objective neural networks [4] (MOBJ-NN) with several selection strategies, and a 'committee machine' aggregating these models.

We have defined the k-probes majority decision predictor as the set of the k top ranked probes together with the majority decision criterion: for any patient case, when the majority of 'pcr' and 'nopcr' predictions of the k top ranked probes was 'pcr', the patient was predicted to be 'PCR', and when the majority was 'nopcr' the patient was predicted to be 'NoPCR'. In case of tie the patient was predicted 'UNSPECIFIED'. When computing the performances of the predictor, a false negative was a PCR patient case predicted NoPCR or UNSPECIFIED, and conversely for the false positives.

Supervised learning involves de minimization of two conflicting objective functions that are related to training set error $\phi_e(\cdot)$ and model complexity $\phi_c(\cdot)$. For most problems, the minima of these two functions do not coincide in the parameter's space, what suggests that they can not be jointly minimized. One of the most popular current approaches for solving this problem [3] aims at minimizing $\phi_c(\cdot)$ by maximizing the separation margin between classes in a kernel-induced feature space. The solution of the optimization problem in the feature space yields a convex optimization formulation, since the kernel non-linear mapping do not appear directly in the final objective functions. Despite of the elegant solution of the convex optimization problem in the feature space for the Support Vector Machine (SVM) [3], the user needs to set a margin parameter in advance. The problem has not changed in its basics and the user is faced again with the same original problem that is to obtain a proper balance between $\phi_e(\cdot)$ and $\phi_c(\cdot)$.

Multi-objective neural networks learning [4] treats neural networks supervised learning as a non-convex optimization problem by trading-off $\phi_e(\cdot)$ and $\phi_c(\cdot)$ directly. In this approach, the concept of global optimality is substituted by the one of Pareto-optimality. After optimization, the Pareto-set contains

the non-dominated solutions [4] that can not be improved in one of the objectives without degrading the other. The decision making procedures follow the Pareto-set generation: a solution is now selected according to a pre-established criteria. The simplest selection approach is to minimize the error of a validation set, but this requires data availability and further computational costs due to cross-validation. Other selection strategies that explore Pareto set properties have been applied successfully to classification and regression problems [5, 6].

5 Results

The performances of the predictors (accuracy, sensitivity and specificity) have been evaluated on the learning set of patient cases and on the independant validation set.

On the validation set, the performances of majority decision predictor were: accuracy=0.86, sensitivity=0.92, specificity=0.84. It outperformed the predictor DLDA-30 [1] whose performances on the validation set were: accuracy=0.76, sensitivity=0.92 and specificity=0.71. On the training set, the majority decision predictor's performances were: accuracy=0.84, sensitivity=0.81, specificity=0.85. We have computed an estimation of the generalisation error by a leave one out cross validation on the set of all the patient cases (82 learning cases and 51 validation cases). The performances were: accuracy=0.77, sensitivity=0.82, specificity=0.75.

The MOBJ solutions were all obtained for a neural network with 10 neurons in the hidden layer. Three different decision strategies that do not depend on a validation set were used for choosing the MOBJ-NN solution. The first one selected the model closer to the point $(0, 1)$ (maximum sensitivity and specificity) in the ROC space [7] given by the training data. The second one selected the model in the intersection between the error and the norm curves given by solutions in the Pareto set. The third one selected the solution with maximum margin in the Pareto set. Data was used in its original expression levels form as well as in its discrete form. Resampling of the minority class was also accomplished in order to compensate the unbalance of the training set. Therefore, models were selected with both balanced and unbalanced training sets. Since the validation set was not used to actually select a model, it will be referred to simply as *test set* from now on.

In order to present the results of all models (including the majority decision predictor) in a single figure, the distance of each solution in relation to the optimal point $(0, 1)$ in the ROC spaces [7] of both training and test data sets were calculated and presented in Figure 5. This form of representing the solutions gives an idea of the sensitivity and specificity performances of each model in both training and data sets. The straight line given by $d_{test} = d_{training}$ corresponds to the solutions that yield the same performance in both training and test sets. The solutions that are below the line perform better in the test set and those that are above the line perform better in the training set. According to this performance measure, the best solutions are those that are close to the line and

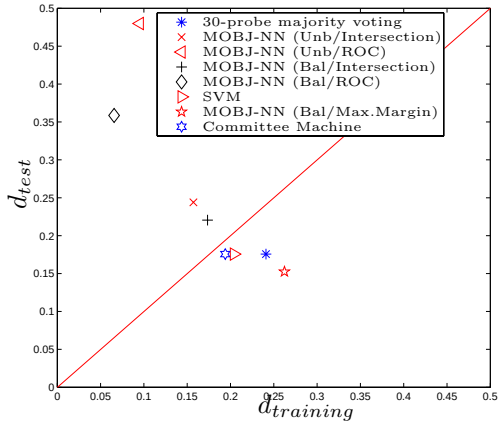

Fig. 1: Distances in the ROC space of the training and test sets for the MOBJ-NN, SVM, 30-probe majority voting models and committee machine.

near the origin of the coordinate system. This assumption is made because the induction principle on which supervised learning is based assumes that training and test data are i.i.d (identically and independently distributed) and, therefore, a model should perform nearly the same in both sets, since they are assumed to have been sampled from the same generator function. Shifts in performance in relation to one of them may be due to a small sample size that may yield a bias in relation to a specific region of the input space. Figure 5 gives, therefore, an idea of which models obey this general trade-off principle between training and test data.

It can be observed from the figure that the MOBJ-NN models with decision strategy based on the ROC curve tend to benefit training regardless of the test set. In fact these solutions tend to minimize the training set error and, consequently, to over-fit the data due to the small number of samples available. The intersection decision making with both balanced and unbalanced training sets resulted on relatively good and well balanced solutions between training and test sets, since they are generally close to the straight line and not far from the origin. The maximum margin MOBJ-NN, SVM and the 30-probe majority voting tended to perform better in the test set, although SVM resulted on a better balance between the two sets. Since there are three models below (test set oriented) and two above (training set oriented) the straight line, a committee machine with two representatives from each side (the closest ones) was also implemented.

The resulting aggregated model yielded a better performance on training while maintaining the best performance in the test set.

6 Conclusion

The method of probes selection that we have proposed has brought predictors which significantly outperformed the best predictor to date for the same data. The simplest of these predictors was the discrete majority decision predictor (DMP). We have shown that under several strategies of solution selection, the MOBJ-NN predictors achieved the same performances on the validation set. None of the strategies of solution selection did bring predictors achieving better performances on the testing set, together with the best performance achieved by the DMP on the validation set. Faced to this tradeoff, we have assessed the performances of several models in the ROC space of the training and validation sets, and we have proposed a 'committee model' for improving the performances of the prediction on the test sets, without decreasing the performances on the validation set.

From this study, we might conclude that a generator function exists which associates the expression levels of the DNA probes to the patient's pathological responses, and that this generator function was, at least partially, identified by the method of probes selection and the different classifier models described in this paper.

Acknowledgements

The authors would like to thank CAPES, COFECUB and CNPq for the support.

References

[1] KR Hess, K Anderson, WF Symmans, V Valero, N Ibrahim, JA Mejia, D Booser, RL Theriault, AU Buzdar, PJ Dempsey, R Rouzier, N Sneige, JS Ross, T Vidaurre, HL Gomez, GN Hortobagyi, and L Pusztai. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of Clinical Oncology*, 24(26):4236–4244, 2006.

[2] R. Natowicz, R. Incitti, B. Charles, P. Guinot, E. G. Horta, L. Pusztai, and R. Rouzier. Prediction of the outcome of preoperative chemotherapy in breast cancer by dna probes that convey information on both complete and non complete responses. *BMC Bioinformatics*, 2008. Accepted.

[3] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[4] R. A. Teixeira, A. P. Braga, R. H. C. Takahashi, , and R. R. Saldanha. Improving generalization of mlps with multi-objective optimization. *Neurocomputing*, (35):189–194, 2000.

[5] T. Medeiros and A. P. Braga. A new decision strategy in multi-objective training of artificial neural networks. In *European Symposium on Neural Networks (ESANN07)*, pages 555–560, 2007.

[6] I. Kokshenev and A. P. Braga. Complexity bounds for radial basis functions and multi-objective learning. In *European Symposium on Neural Networks (ESANN07)*, pages 73–78, 2007.

[7] W.S. Andrus and K.T. Bird. Radiology and receiver operating characteristic (roc) curve. *CHEST*, 67(4):378–379, 1975.