Anais do IX Congresso Brasileiro de Redes Neurais / Inteligência Computacional (IX CBRN)
Ouro Preto 25-28 de Outubro de 2009 © Sociedade Brasileira de Redes Neurais

SEMI-SUPERVISED MODEL APPLIED TO THE PREDICTION OF THE RESPONSE TO PREOPERATIVE CHEMOTHERAPY FOR BREAST CANCER

FREDERICO COELHO[*], ANTÔNIO DE PÁDUA BRAGA[*], RENÉ NATOWICZ[†], ROMAN ROUZIER[‡]

[*] *Universidade Federal de Minas Gerais*
CPDEE - Brazil

[†] *Université Paris-Est, ESIEE-Paris*
Département d'ínformatiquex - France

[‡] *Hôpital Tenon, Départment of Gynecology*
Paris, France

Emails: fredgfc@cpdee.ufmg.br, apbraga@ufmg.br, r.natowicz@esiee.fr, roman.rouzier@tnn.aphp.fr

Abstract— Breast cancer is the second most frequent one, and the first one affecting the women. The standard treatments has three main stages: a preoperative chemotherapy followed by a surgery operation, then an post-operatory chemotherapy. Because the response to the preoperative chemotherapy is correlated to a good prognosis, and because the clinical and biological informations do not yield to efficient predictions of the response, a lot of research effort is being devoted to the design of predictors relying on the measurement of genes' expression levels. In the present paper, we report our works for designing genomic predictors of the response to the preoperative chemotherapy, making use of a semi-supervised machine learning approach. The method is based on margin geometric information of patterns of low density areas, computed on a labeled dataset and on an unlabeled one.

Keywords— Transductive learning , Unlabeled data set , Semi-supervised learning.

Resumo— O câncer de mama é o segundo tipo mais freqüente, sendo o primeiro em mulheres. O tratamento padrão possui três fases principais: uma quimeoterapia pré-operatória, seguida por cirúrgia e, em seguida, uma quimeoterapia pós-operatória. Porque a resposta à quimeoterapia pré-operatória está correlacionada com um bom prognóstico, e porque a informação clínica e biológica não levam à previsões eficientes desta resposta, uma grande esforço de investigação está sendo dedicado ao projeto de preditores baseando-se nos níveis de expressão gênica. No presente trabalho, relatamos os nossos trabalhos para a concepção de preditores genômicos da resposta à quimioterapia pré-operatória, fazendo uso de uma máquina de aprendisagem semi-supervisionada. O método baseia-se na informação de margem geométrica dos padrões das áreas de baixa densidade, calculada sobre um conjunto de dados rotulados e sobre outro não rotulado.

Keywords— Aprendizado transdutivo, Dados não rotulados, Aprendizado semi-supervisionado.

1 Introduction

Predicting the response of a patient to preoperative chemotherapy from the measurement of genes expressions is being a main issue in clinical cancer research since DNA microarrays have become available, about ten years ago. The importance of developing such predictors relies on the fact that only 30% of patients have a positive response to the treatment and, in absence of efficient predictors of the response, most of the patients are allocated to the standard treatment. A lot of statistical and machine learning models have been developed to address the problem (Cooper, n.d.; Glas et al., n.d.; Ancona et al., n.d.; Michiels et al., n.d.), but no genomic predictor is yet accurate enough to be used in clinical routine. Among the main issues in the development of such models are: (i) selecting relevant genes to enter the predictors among thousands of genes whose expression levels are measured by DNA microarrays (the vast majority of them being not involved in the response to the chemotherapy treatments), (ii) the small number of cases compared to the numbers of features (genes expressions), (iii) the representativeness of the data. These difficulties are challenging for the development and the validation of prediction models.

In the particular case of the application reported in this article, the dataset is, for each patient case, the expressions of a set of genes considered as relevant markers of the response to the chemotherapy, and the outcome of the treatment. The data themselves has been collected in a clinical trial in which 133 patients were embedded. The clinical trial was jointly conducted at the Institut Gustave Roussy (Villejuif, France) where 51 patients were cared, and at the MD Anderson Cancer Center (Houston, USA), where 82 patients were cared. All the patients were allocated to a preoperative chemotherapy treatment, to which each of them revealed to be either responder (Pathological Complete Response, i.e. the result of the treatment was that tumor had vanished) or non-responder (in the case of a residual disease). Because the data were collected in two far distant area (France and USA), there may be some genetic bias in the outcomes, suggesting that a semi-supervised learning (SSL) approach could be relevant for addressing the problem (Chapelle et al., 2006; Zhu, 2008).

The main idea of SSL is to design the model not only on the basis of a labeled data set (gene expressions and known responses, usually called the training set), but moreover by making use of a structural information obtained from an additional unlabeled set, called the working set. An approximation of the general separating function of the two classes (responders and non

 © Sociedade Brasileira de Redes Neurais

responders) is induced from the training set and from some assumptions about the working set. The basic assumption of our work is that the separating surface should have maximum margins in both the training and working sets.

A new SSL method for artificial neural networks (ANN) learning, based on the geometrical computation of the separation margin is presented and applied to the problem. The separation region is associated to the lowest density region in the input space formed by both data sets. A method to geometrically identify the separation region and to compute the margins is also presented. The final solution is then obtained by addressing the problem as a multi-objective learning problem and by selecting the final model as the one which maximizes the separation margins in both datasets.

In this paper we will review the basic concepts that are discussed or applied in our work, then we will describe the implemented semi-supervised method, we will present the results that we have obtained and compare them to those of other approaches for the same data. Finally we will discuss the results and some issues of this approach.

2 Review

2.1 *Semi-Supervised Learning*

In semi-supervised learning we are concerned by using the information conveyed by the unlabeled data set, in addition to that of the labeled (training) set. The use of the unlabeled data together with a small amount of labeled data can improve the efficiency of the learning process. SSL can be of great value when labeled data is difficult to obtain (Zhu, 2008). In contrast, the acquisition of unlabeled data can be of lesser cost.

In SSL, a *supervisor* can compare the outpouts of the network with the known labels of the learning dataset, and make the necessary adjustments of the network's weights. But the *supervisor* is of no use for the unlabeled data.

2.2 *Sliding Mode of Control for Multi-Objective Neural Networks (SMC-MOBJ).*

A multi-objective algorithm (Costa et al., 2007) aims at reaching a balance between the *bias and variance* of a neural network, by selecting Pareto solutions in the objectives space defined by the vector norm $||w||$ of the network's weights, and the classification error, e, on the training set. The sliding mode algorithm is capable of generating arbitrary trajectories in the space of the objectives and reaching any solution $(\mathbf{e}_k, \|\mathbf{w}_k\|)$ in the space of the objectives defined by the sum of the quadratic error $\mathbf{e}_k$ and the weights vector norm $\|\mathbf{w}_k\|$. This multiobjective algorithm minimizes two sliding modes surfaces that are defined by $S_v = (\mathbf{e} - \mathbf{e}_k)$ and by $S_{\|w\|} = \left(\|w\|^2 - \|w_k\|^2\right)$ (Costa et al., 2007).

2.3 *Margin*

The margin is defined as the distance between the separation hyperplane and the closest points of the data set (Haykin, 2001). For support vector machines (SVM) these points are called support vectors. The optimal separation hyperplane is the one equidistant to both classes. Figure 1 is an example of the concept of optimal hyperplane and margin, denoted by ρ in the figure. The equations to compute the margin ρ are given by equation (1), (2), and (3) (Shawe-taylor and Cristianini, n.d.) :

$$\rho = \sum_{i=1}^{m} y_i d_i \tag{1}$$

$$d_i\left(\mathbf{w}, b, x\right) = \frac{\left(\mathbf{w} \cdot x_i + b\right)}{\|w\|} \tag{2}$$

$$\rho\left(\mathbf{w}, b\right) = \frac{2}{\|\mathbf{w}\|} \tag{3}$$

where m is the number of selected patterns for computing the margin, and d is the distance between the pattern i and the separation hyperplane (computed according to equation 2.) In general, the margin is computed for the support vectors because they are in each class, the points which are the closest to the separation hyperplane. These points synthesize all the information about the classes that define the separation hyperplane. In other words, training the network with all the input patterns is equivalent to training it with the only support vectors. The optimal hyperplane is found by maximizing the margin $\rho\left(\mathbf{w}, b\right)$ subject to $y_i\left[\left(\mathbf{w} \cdot x_i\right) + b\right] \geq 1$, (Gunn, 1997), which takes the form of equation 3.

Maximizing the separation margin is equivalent to minimizing the weight vectors of the network ($\mathbf{w}$) (Haykin,

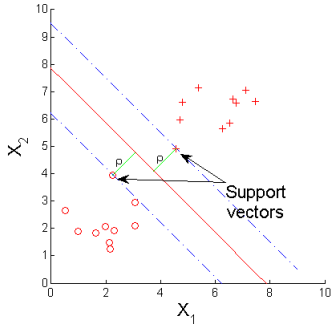
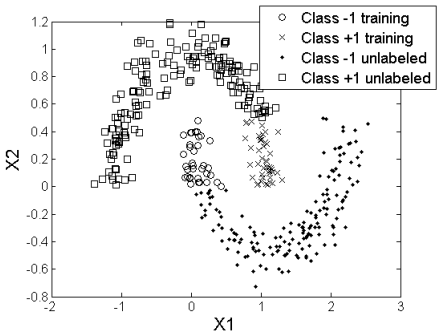
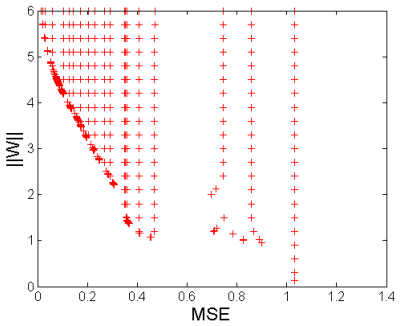


Figure 1: Illustration of an optimal hyperplane for linearly separable patterns.

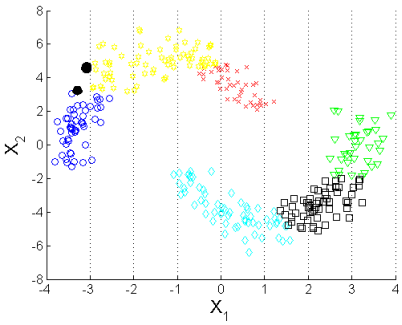


(a) The two moons problem

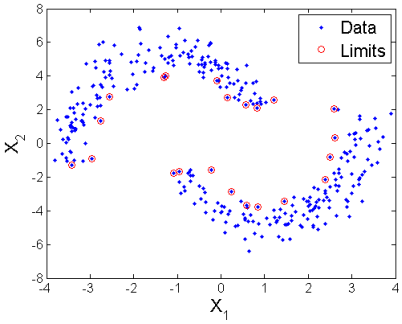


(b) Solution grid in the objectives space

Figure 2: The two moons problem and its solution grid



(a) Slicing method applied to the two moons problem



(b) Patterns selected in all simulation (100%)

Figure 3: MILFAT method

2001). Given a set of solutions we can compute geometrically the distance between the separation hyperplane and a given data point, then select the solution of maximum margin.

3 Description of the new method

Given a set of solutions (fig. 2(b)) obtained from a training dataset generated by the SMC-MOBJ method (Costa et al., 2007), we take into consideration the distribution of both patterns of the labeled and unlabeled sets in order to select the solution of best generalization.

This solution is located between the classes, in the center of the area of patterns lower density, considering that such an area exists. This optimal separation solution has maximum margins, and we can compute it geometrically in relation to the patterns that deleneate the lower density area between the classes in all its extensions (Coelho, 2009). in this section we will illustrate the method on a well known two dimensional problem, the *two moons problem* (figure 2(a)). The figure 2(a) depicts the patterns belonging to the labeled (training) and unlabeled sets.

3.1 First proposition: Method for Identifying the Borders by Slicing - MILFAT

The method for identifying the limits by slicing (MILFAT) is based on a grouping method (Fuzzy c-means clustering – FCM) (Bezdec, 1981). We have applied the FCM method to the unlabeled data set for any final number of partitions c. In a second stage, for each axis of the input space, we have separated the data into bands and later we have determined the borders between the regions of the FCM partition, 'walking' along the sorted data values of the others axis (fig. 3(a)). Most of these borders were real frontiers between classes. However, some meaningless borders (artifacts) were also present. This process was iterated for each axis.

To discard the artifact borders, the FCM algorithm was run for an additional iteration on the unlabeled set, with a number of partitions different from the one used in the first execution of the algorithm. We have compared the borders found in the two simulations of the FCM and we have checked them against the one that had been selected in all the simulations. Those of the patterns that were next to the regions of lower density between classes tended to be selected in all the simulations with different partitions, while the artifacts (those which were on the borders between partitions but in the same class) tended to be differ from one execution to another (cf. fig. 3(b)).

3.2 Second proposition: Method for Identifying the Borders by Probabilities - MILP

In this method, at the end of the FCM simulation, instead of slicing we have used the information of the FCM resultant probability matrix, taking as the borders the n patterns of lowest probability among all of those that had been clustered in one partition c (and this, for each partition). Iterating this process for different numbers of FCM partitions, we have marked as border patterns those chosen in all the simulations.

3.3 The Geometric Margin Computation

For computing the margins, we need to go back to equation 1. Our proposition is to compute the distance between the separation hyperplane and the selected input patterns *through all the borders* between the classes in the hidden layer of the network. For reasons that are discussed in (Coelho, 2009) we have defined three different ways to compute the geometric margin ($\rho_S = \rho_1 + \rho_2$, $\rho_M = \rho_1 * \rho_2$ and $\rho_D = min(\rho_1, \rho_2)/max(\rho_1, \rho_2)$) to evaluate which of them was the best one. In these expressions, ρ_1 is the sum of the distances of the limit patterns of class -1 to the hyperplane, weighed by the network output, and ρ_2 is the same for class +1, as in equation 1.

3.4 Selection of the Best Solution

We have generated a solution set in the objectives space by training the network with the inductive set (fig. 2(b)) using a sliding mode MOBJ algorithm (Costa et al., 2007). For each solution we have summed the computed geometric margin for the inductive set with the one computed for the unlabeled set, then we have chosen the solution that had the largest total margin. We have chosen the grid solution $\mathbf{G} : (e^*, \mathbf{w}^*)$ that maximized the total geometric margin ρ_{tot} explicited in (4). The values $\rho_{labeled}$ and $\rho_{unlabeled}$ are the results defined in section 3.3 for labeled and unlabeled sets respectively.

$$\rho tot = \rho_{labeled} + \rho_{unlabeled} \quad (4)$$

4 Applying the Method to Genomic Datasets

The data of this application of the method to the prediction of the response to preoperative chemotherapy for breast cancer are, on the one hand, the expression levels of a set of genes measured on tumor tissues for 133 patients. The patients with no residual disease at the time of surgery were responders to the treatment (pathological complete response – PCR –), while those with residual disease were non responders (NOPCR patient cases). Our goal was to design a model relying on the gene expression levels for predicting the outcome of the treatment (for patient cases that were not member of the learning set).

The 82 patient data from the MD Anderson Cancer Center (Texas, USA) were the training dataset, and the 51 patient data from the Institut Gustave Roussy (Villejuif, France) were the unlabeled dataset. The works of Horta (Horta, 2008), Hess (K.R. Hess and Pusztai., 2006), Natowicz (Rene Natowicz and Rouzier., march 2008) and Braga (R. Natowicz and Costa, 2008) (Braga et al., n.d.), have made use of these data to select relevant probes for prediction. Among them, we have selected three probes sets for our application of our semi-supervised learning method : the 30 probes set selected by Natowicz (Rene Natowicz and Rouzier., march 2008), the 18 probes set and the 11 probes set selected by Horta (Horta, 2008)

The results (table 1) found by the SSL method based on geometric margin are very interesting, because the results obtained at works like Horta's one (Horta, 2008), (table 2), were achieved considering all training data as labeled ones. The SSL method consider labeled data and uses aditional information from unlabeled data distribution. In tables **FP**, **FN**, **Ac**, **Se** and **Sp** means, respectively, false positive results, false negative, acuracy (%), sensitivity (%) and specificity (%) for both training and validation sets. Model A refers to set of 18 probes with MILP/MILFAT methods for $\rho_S/\rho_M/\rho_D$ methods calculation, model B is 11 probes with MILP/MILFAT for $\rho_S/\rho_M/\rho_D$, model C is 30 probes with MILP for ρ_S/ρ_M, model D is 30 probes with MILP for ρ_D, model E is 30 probes with MILFAT for ρ_S/ρ_M and model F is 30 probes with MILFAT for ρ_D. In table 2 model G is 18 probes with SVM RBF, model H is 32 probes with SVM RBF and model I is 18 probes with LASSO.

Table 1: MSMG's results

	Training					Validation				
Models	FP	FN	Ac	Se	Sp	FP	FN	Ac	Se	Sp
A	2	3	93,9	85,7	96,7	7	2	82,4	84,6	81,6
B	3	3	92,7	85,7	95,1	8	2	80,4	84,6	78,9
C	11	2	84,1	90,5	82,0	9	1	80,4	92,3	76,3
D	9	3	85,4	85,7	85,2	10	1	78,4	92,3	73,7
E	13	3	80,5	85,7	78,7	8	1	82,4	92,3	78,9
F	12	3	81,7	85,7	80,3	8	1	82,4	92,3	78,9

Table 2: Horta's results

	Training					Validation				
Modelo	FP	FN	Ac	Se	Sp	FP	FN	Ac	Se	Sp
G	1	3	95,1	85,7	98,4	5	2	86,3	84,6	86,8
H	2	3	93,9	85,7	96,7	6	1	86,3	92,3	84,2
I	2	4	92,7	81,0	96,7	4	3	86,3	76,9	89,5

The best SSL method results are comparable to best results achieved in (R. Natowicz, 2009) that used a SVM with linear kernels (See table 4 of (R. Natowicz, 2009)). Note that we achieved acuracy values for training and validation a little smaller than the best result on Horta's work, however, it is still a good result. We get the same sensibilities values and ours specificities were slightly worse compared to the best result in table 2.

5 Discussion

The semi-supervised method based on geometric margin was applied to the problem of predicting the response to preoperative chemotherapy in the treatment of breast cancer. The results obtained with 18 and 11 probes are comparable to the best results achieved by Horta (Horta, 2008). The results are balanced, stable, with low error rate. The results obtained by computing the geometric margins in different ways were very close for the same probes set (defining the best way to compute it is an important issue for the SSL model (Coelho, 2009)). MILFAT method supposes to define a number of slices for searching and, as the algorithm has high computational cost, this factor is of extreme importance: more slices, higher will be processing time. In MILP, one has to set a cutoff point of the probability of a pattern belonging to the limit of the partition. The number of times the FCM has to be run in both methods and the number of partitions on each of them can also influence the outcome. However these parameters appeared to have a lesser influence on the performances of the predictor.

References

Ancona, N., Maglietta, R., Piepoli, A., D'Addabbo, A., Cotugno, R., Savino, M., Liuni, S., Carella, M., Pesole, G. and Perri, F. (n.d.). On the statistical assessment of classifiers using dna microarray data.

Bezdec, J. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York.

Braga, A. P., Horta, E. G., Natowicz, R., Rouzier, R., Incitti, R., Rodrigues, T. S., Costa, M. A., Pataro, C. D. M. and Çela., A. (n.d.). Bayesian classifiers for predicting the outcome of breast cancer preoperative chemotherapy., *In The Third International Workshop on Artificial Neural Networks in Pattern Recognition (ANNPR08)* .

Chapelle, O., Schölkopf, B. and Zien, A. (eds) (2006). *Semi-Supervised Learning*, MIT Press, Cambridge, MA.
URL: *1*

Coelho, F. G. F. (2009). *Modelo semi-supervisionado aplicado à previsão da eficiência da quimioterapia neoadjuvante no tratamento de câncer de mama*, Dissertação de mestrado, Universidade Federal de Minas Gerais.

Cooper, C. (n.d.). Applications of microarray technology in breast cancer research.

Costa, M. A., de Pádua Braga, A. and de Menezes, B. R. (2007). Improving generalization of mlps with sliding mode control and the levenberg-marquardt algorithm, *Neurocomput.* **70**(7-9): 1342–1347.

Glas, A., Floore, A., Delahaye, L., Witteveen, A., Pover, R., Bakx, N., Lahti-Domenici, J., Bruinsma, T., Warmoes, M., Bernards, R., Wessels, L. and Veer, L. V. (n.d.). Converting a breast cancer microarray signature into a high-throughput diagnostic test.

Gunn, S. (1997). Support vector machines for classification and regression, *Image Speech & Intelligent Systems Group - University of southampton* .

Haykin, S. (2001). *Redes Neurais: Princípios e Prática*, Bookman.

Horta, E. G. (2008). *Previsores para a eficiência da quimioterapia neoadjuvante no câncer de mama*, Dissertação de mestrado, Universidade Federal de Minas Gerais.

K.R. Hess, K. Anderson, W. S. V. V. N. I. J. M. D. B. R. T. A. B. P. D. R. R. N. S. J. R. T. V. H. G. G. H. and Pusztai., L. (2006). Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer., *Journal of Clinical Oncology,24(26):42364244* .

Michiels, S., Koscielny, S. and Hill, C. (n.d.). Prediction of cancer outcome with microarrays: A multiple random validation strategy.

R. Natowicz, A. P. Braga, R. I. E. G. H. R. R. T. S. R. and Costa, M. A. (2008). A new method of dna probes selection and its use with multi-objective neural networks for predicting the outcome of breast cancer preoperative chemotherapy., *European Symposium on Artificial Neural Networks* .

R. Natowicz, R. Incitti, R. R. A. C. A. B. E. H. T. R. M. C. C. P. (2009). Downsizing multigenic predictors of the response to preoperative chemotherapy in breast cancer, *Investigating Human Cancer with Computational Intelligence Techniques, KES International Recent Research Results Series* pp. 29–41.

Rene Natowicz, Roberto Incitti, E. G. H. B. C. P. G. K. Y. C. C. F. A. L. P. and Rouzier., R. (march 2008). 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, 9:149* .

Shawe-taylor, J. and Cristianini, N. (n.d.). Smola, bartlett, scholkopf, and schuurmans: Advances in large margin classifiers, introduction to large margin classifiers.

Zhu, X. (2008). Semi-supervised learning literature survey.