Early Breast Cancer Prognosis Prediction and Rule Extraction Using a New Constructive Neural Network Algorithm

Leonardo Franco[1], José Luis Subirats[1], Ignacio Molina[2], Emilio Alba[3], and José M. Jerez[1]

[1] Departamento de Lenguajes y Ciencias de la Computación
Escuela Técnica Superior de Ingeniería en Informática
Universidad de Málaga,
Campus de Teatinos S/N, 29071 Málaga, Spain
[2] Departamento de Tecnología Electrónica
Escuela Técnica Superior de Ingeniería en Telecomunicación
Universidad de Málaga,
Campus de Teatinos S/N, 29071 Málaga, Spain
[3] Servicio de Oncología,
Hospital Clínico Universitario Virgen de la Victoria, 29071, Málaga, Spain

Abstract. Breast cancer relapse prediction is an important step in the complex decision-making process of deciding the type of treatment to be applied to patients after surgery. Some non-linear models, like neural networks, have been successfully applied to this task but they suffer from the problem of extracting the underlying rules, and knowing how the methods operate can help to a better understanding of the cancer relapse problem. A recently introduced constructive algorithm (DASG) that creates compact neural network architectures is applied to a dataset of early breast cancer patients with the aim of testing the predictive ability of the new method. The DASG method works with Boolean input data and for that reason a transformation procedure was applied to the original data. The degradation in the predictive performance due to the transformation of the data is also analyzed using the new method and other standard algorithms.

1 Introduction

The evaluation of the probability of cancer relapse for breast cancer patients that underwent surgery is crucial as it influences much the choice of the adjuvant therapy. Adjuvant treatments reduce the risk of relapse but can also harm patients' health and have secondary non-desirable effects. Thus, knowing the benefits of applying in individual cases a certain treatment helps on the decision-making process. Regression models are the standard tools for survival analysis but in recent years the use of alternative models, like neural networks, have much increased due to their improved performance [1,2]. Nevertheless, the application and analysis of neural networks suffer from two main problems: first,

F. Sandoval et al. (Eds.): IWANN 2007, LNCS 4507, pp. 1004–1011, 2007.

a key step in applying a neural network to a problem is the selection of the architecture and the value of the parameters involved, and second the fact that understanding the underlying rules of the predictive process are quite complex for neural networks. Recently, in Ref. [3] we have developed a new constructive neural network algorithm (DASG) that permits a relatively easy rule extraction procedure. Constructive neural networks algorithms work by analyzing the set of examples of the problem and adding iteratively new neurons when is needed. The recently introduced DASG algorithm works by decomposing the input function in simpler ones according to the most complex variable, has the advantage of having no tuning parameters and as works with AND or OR output neurons the rule extraction procedure can be easily applied. One problem at the time of the implementation of the DASG algorithm is that it works with binary input data, implying that the original real-input or categorical data of the problem should be transformed into Boolean variables. The transformation of variables is also of interest from a practical point of view as quite often the protocol for applying medical treatments uses indexes for which a categorization of the original values is needed. In this sense, one issue of interest is to quantify how much information, and consequently prediction accuracy, is lost when the data is transformed. Another issue with real data, and in particular with clinical data is that often is unbalanced, i.e., there exists more negative than positive examples. We analyze in this work the effect of artificially balancing the cases in the prognosis accuracy of the predictive methods used. We compare the predictive performance of the new DASG algorithm with the obtained from standard feed-forward neural networks (FFNN) and decision trees models.

2 Methods

2.1 The DASG Algorithm and the Rule Extraction Procedure

The DASG algorithm belongs to the class of constructive methods and was designed to create compact neural network architectures for Boolean input problems. The algorithm decomposes the original problem by dividing it into subproblems of reduced complexity. The splitting is carried out using the variable with the highest influence, considered as the most complex or conflictive, among the non-unate variables. The influence of a Boolean variable is defined as the number of times that a change of the variable from 0 to 1 produce a change on the output of the function. A function is unate if all its variables are unate; and a variable is unate if its influence is only positive or negative but not both at the same time (a variable that has both positive and negative influences is called binate). Furthermore, unateness is related to linear separability, as it is known that all linear separable functions are unate. The splitting procedure, that simplifies function complexity, is repeated until the obtained function is linearly separable, in a way that each time the problem is split according to a non-unate variable, a neuron is added to the single hidden neuron of the architecture. The DASG algorithm is a parameter free method and the only choice to be done for the application of the algorithm is to decide whether to use an AND or OR

representation. The output neuron of the whole architecture has to be chosen at the beginning of the process among an AND or OR functions. The use of AND or OR functions as output and the general small size of the created networks permit a relatively simple rule extraction procedure as the rules created can be analyzed directly from the function that the hidden nodes neurons compute. The DASG algorithm has been tested in the construction of efficient logical circuits and also its generalization ability tested with these logical problems with very satisfactory results.

In this study, the performance of the DASG algorithm was compared with the obtained from other standard algorithms for classification tasks, like standard feedforward neural networks and decision trees. The neural network and decision trees models were implemented using the open source platform [4]. Some optimization on the parameters used for these two algorithms was tried but the best results were obtained using the default parameters. The algorithm C4.5 was used for the decision tree model [5], as this is a kind of standard method, often used in comparisons. To quantify the performances of the methods, the generalization ability was measured together with the balanced generalization ability defined as the average between the generalization observed for positive and negative cases. Another estimator of the performance of the methods that is often used and preferred in medical prognosis is the area under the ROC plot [6,7,8], that we also compute for the different methods.

2.2 The Breast Cancer Dataset and Its Transformation into Boolean Variables

Data were collected from the El Alamo 1 project, the largest database on breast cancer in Spain. The dataset analyzed in this study includes demographics, therapeutic and recurrence-survival information from 1960 women patients with operable invasive breast cancer diagnosed in 32 different hospitals belonging to the Spanish Breast Cancer Research Group (GEICAM) between the years 1990 and 1993. This study used the set of clinical and pathological variables selected in [2] as more significant prognostic factors in the prediction of patients outcome. The analysis was restricted to estimate early breast cancer relapse and then the time was not a relevant variable. The state of the patient used as output target for the methods was a binary variable indicating whether the patient has relapsed or not before the 30th month after surgery. The period of time after surgery and before the 30th month is a high risk one, in which a peak in the hazard function can be clearly identified [9,2], and thus is important to analyze the factors that lead to this early relapse. The DASG algorithm works with Boolean input data and as the original breast cancer dataset contained real and categorical data, the data was transformed into binary variables. The variables Age, Tumor size, Number of axillary lymph nodes, originally represented by real values, were each coded by two binary variables representing 4 categories. The choice of using 2 Boolean variables is justified by the fact of trying to keep the total number of input variables low, minimizing the computational times and to facilitate the rule extraction procedure. As the reduction of performance due to the binarization

procedure was not dramatic we decided to keep this choice and not try a larger number of categories. The 4 categories were chosen by dividing the original range of the variables in 4 intervals trying to allocate the same number of cases on each of the intervals. In table 1, the range , number of values and binary variables used are indicated for the 3 original real-value covariates. The dataset includes other two categorical variables for the type of the tumor (3 cases) and the treatment applied (8 cases). For these two covariates a binary code was also used, needing 2 variables for the tumor type and 3 for the type of treatment.

Table 1. The original dataset of breast cancer patients contained real input values for the covariates Age, Size of the tumor and number of axillary nymph nodes and they were transformed into binary variables. Each covariate range of values were divided in 4 groups trying to allocate in them a similar number of cases. The variables Histological grade and Type of treatment were originally categorical variables and they were coded as binary variables.

Range	Size	Category
Age		$x_0\ x_1$
$[25-47)$	492	00
$[47-57)$	508	01
$[57-66)$	462	10
$[66-90]$	498	11
Size		$x_2\ x_3$
$[0.2-2.0)$	480	00
$[2.0-2.5)$	560	01
$[2.5-4.0)$	442	10
$[4.0-13.0]$	478	11
#-nodes		$x_4\ x_5$
$[0.0-0.2)$	482	00
$[2.0-2.5)$	458	01
$[2.5-3.6)$	542	10
$[3.6-35]$	478	11
Hist. grade		$x_6\ x_7$
1	381	00
2	1106	01
3	473	10
Treatment		$x_8\ x_9\ x_{10}$
Radiotherapy	102	000
Hormonetherapy	458	001
Chemotherapy	225	010
Hormone-radio.	400	011
Chemo-radio	250	100
Chemo-hormone	208	101
Radio-hormone-chemo	206	110
No treatment	112	111

3 Results

3.1 The Predictive Accuracy of the Methods and the Effect of Balancing the Dataset

We first analyzed the effect of the variable transformation process applied to the input data in the predictive accuracy using as predictive methods a standard feed-forward neural network model and a decision tree. The methods were implemented using the open source platform Weka [4] with the default parameter setting. A 10-fold cross validation procedure was applied to the 1960 records of patients, testing the generalization ability of the methods, calculating the ROC area and computing as well the balanced generalization defined as the average between the generalization ability for positive and negative cases. The results indicate a degradation of a 8% in the predictive accuracy due to the data transformation for the case of FFNN (A reduction in the ROC area from 0.72 to 0.66). The DASG algorithm can only be tested with the categorized variables and the performance obtained with the new method was quite similar to the obtained by the C4.5 decision tree method. The results obtained with the 3 methods used are indicated in Table 2 and Table 3. The results in Table 2 correspond to the application of the 3 different algorithms to the original unbalanced dataset, while those shown in Table 3 correspond to the balanced dataset. The best results in terms of the area ROC (0.72) were obtained for the FFNN model using the original unbalanced dataset. Table 3 does not include the results from the application of the DASG algorithm, as the balancing procedure does not affect its operation and thus the results are exactly the same as those obtained in Table 2.

Table 2. Performance measures of the prediction accuracy of the early breast cancer relapse using the original supplied dataset and a categorized one obtained with three different methods: Decision trees (C4.5), Feed-forward neural networks (FFNN) and DASG. A 10-fold cross validation procedure was applied to the set of 1960 records of patients. The Area under the ROC curve, the generalization ability and the balanced generalization ability are shown for the different methods utilized.

	ROC Area		Generalization ability		Balanced Accuracy	
Method	Orig. Data	Categ. Data	Orig. Data	Categ. Data	Orig. Data	Categ. Data
C4.5	0.62	0.50	82.55	82.96	54.06	50.00
FFNN	0.72	0.66	83.01	80.81	55.18	57.57
DASG	–	0.60	–	79.33	–	59.72

The original data was significantly unbalanced, containing 5 times more negative examples than positive ones. Thus, we decided to artificially balance the dataset by replicating the positive cases in the training set. The results using this balanced dataset show that in terms of the ROC area, the preferred estimator for the accuracy of the predictions, there are not much changes and the obtained values remain similar, as it can be seen from the results shown in

Table 3. Performance measures of the prediction accuracy of the early breast cancer relapse using the balanced categorized dataset with three different methods: Decision trees (C4.5), Feed-forward perceptron (FFNN) and DASG. The Area under the ROC curve, the generalization ability and the weighted generalization ability are shown for the different methods utilized.

	ROC Area		Generalization ability		Balanced Accuracy	
Method	Orig. Data	Categ. Data	Orig. Data	Categ. Data	Orig. Data	Categ. Data
C4.5	0.60	0.61	73.31	65.71	60.49	60.87
FFNN	0.71	0.65	68.42	63.47	66.56	60.79

Table 4. Set of rules obtained for the 5 most relevant neurons in the hidden layer of an architecture generated by the DASG algorithm

$$R1(x_0, x_1, x_2, x_4, x_8, x_9, x_{10}) = \bar{x}_0\bar{x}_1x_8\bar{x}_9\bar{x}_{10} + x_0\bar{x}_1\bar{x}_2x_4x_8(x_9\bar{x}_{10} + \bar{x}_9x_{10})$$
$$R2(x_0, x_1, x_2, x_3, x_9, x_{10}) = x_0x_1\bar{x}_3[x_2(1 + x_9x_{10}) + \bar{x}_2\bar{x}_9x_{10}]$$
$$R3(x_0, x_2, x_5, x_6, x_8, x_9, x_{10}) = \bar{x}_0\bar{x}_2x_5[x_8x_9\bar{x}_{10} + \bar{x}_6(x_9x_{10} + \bar{x}_9\bar{x}_{10})]$$
$$R4(x_1, x_2, x_3, x_7, x_8, x_9, x_{10}) = \bar{x}_1\bar{x}_2\bar{x}_3x_7x_8\bar{x}_9\bar{x}_{10}$$
$$R5(x_2, x_5, x_6, x_7, x_8, x_9, x_{10}) = \bar{x}_7\bar{x}_8\bar{x}_{10}(x_2x_5\bar{x}_6x_9 + \bar{x}_2\bar{x}_5x_6\bar{x}_9)$$

Table 3. Only for the case of the decision tree algorithm the ROC area improves much for the categorized data after balancing the cases (AUROC=0.61) in comparison to the original dataset when the C4.5 algorithm have a much lower performance (AUROC=0.5).

3.2 Rules Obtained by the DASG Algorithm for the Early Breast Cancer Relapse Problem

The neural architectures generated by the DASG algorithm have an output neuron that can be selected between an AND or OR function. As the dataset considered have a number of negative cases lower than the number of positive ones, we decided to work with an OR output function, because in this case the function obtained by the algorithm will contain more positive cases (in comparison to the solution corresponding to the AND output case) and this can be interpreted as giving more importance to the positive cases. For the rule extraction procedure, we analyzed directly the functions obtained for the hidden neurons, as whenever a hidden neurons is active, the OR output neuron will be also active. We show in table 4 the rules obtained by the 5 more relevant neurons (those that gets activated more often by the training dataset) of a DASG generated architecture for a single dataset (the first one in the cross validation procedure) for which the ROC area was 0.62.

Aa an example of the interpretation of the rules, we consider the case of Rule 1 (indicated in Table 4 by $R1$). According to the codification of the input variables indicated in Table 1 third column under "Category", this rule indicates that a group of patients with ages between 25-47, that have been treated with chemo plus radio therapies have a large probability of relapse (first part of Rule 1 : $\bar{x}_0\bar{x}_1x_8\bar{x}_9\bar{x}_{10}$). The second part of Rule 1 ($x_0\bar{x}_1\bar{x}_2x_4x_8(x_9\bar{x}_{10} + \bar{x}_9x_{10})$) suggests that patients in the group age 57-66 with a tumor size between 0.2 and 2.5, with a number of nodes larger than 2.5, that have been treated with chemo plus hormone therapies or treated by radio plus hormone plus chemo therapies have a large probability of relapse.

4 Discussion

Using a recently introduced method (DASG algorithm) we have analyzed the prediction accuracy and the rules obtained using a large dataset of breast cancer patients for which the probability of cancer relapse has to be computed. The prediction accuracy of the new method was analyzed in comparison to standard algorithms like decision trees (C4.5 algorithm) and standard feedforward neural networks (FFNN) trained by backpropagation. The new implemented method works with binary input data and the results using the categorized dataset shows a performance of the algorithm a little bit lower (AUROC=0.60) in comparison to the obtained with the FFNN (AUROC=0.66), but similar to the obtained with the C4.5 algorithm with balanced data (AUROC=0.61). It is worth noting that the effect of balancing the data affected the C4.5 algorithm when the categorized data was used, increasing its AUROC from 0.5 to 0.61.

Regarding the comparison of the performance prediction using the original dataset and the categorized one, for the case FFNN the value of 0.72 obtained for the ROC area with the original dataset was reduced to 0.62 when the categorized dataset was used. We have applied a simple procedure, grouping a similar number of cases in 4 different classes. We are aware that better strategies can be applied and we will test them in the near future.

We have also analyzed the effect of balancing the number of positive and negative cases on the predictive accuracy of FFNN and C4.5 algorithms and obtained that there was not much differences regarding the AUROC curve when the original (non-balanced) dataset was considered.

The DASG algorithm permits a simple extraction of the rules created by the algorithm and we have shown that the procedure can be implemented in a real task like the breast cancer prognosis problem. In this work , we have just limit our study to show that the extraction of the rules is straightforward but have not analyzed the obtained rules as this involves a much detailed and careful analysis.

As a conclusion, in this first application of the new DASG algorithm we have shown that the method can be applied to real data and that the rules governing the predictions can be extracted without much effort. A number of improvements can be done to the present work, and we are currently working on them, in particular on the real-to-Boolean variable transformation and on the interpretation and analysis of the obtained rules.

Acknowledgements

The authors acknowledge support from CICYT (Spain) through grant TIN2005-02984 (including FEDER funds). The authors acknowledge the GEICAM group for granting permission of use for the "El Alamo" dataset. Leonardo Franco acknowledges support from the Spanish Ministry of Education and Science through a Ramón y Cajal fellowship.

References

1. Biganzoli, E., Boracchi, P., Coradini, D., Daidone, M.E, Marubini, E.: Prognosis in Node-negative Primary Breast Cancer: a neural network analysis of risk profiles using routinely assessed factors. Annals of Oncology 14, 1484–1493 (2003)
2. Jerez, J.M., Franco, L.E., Alba, E., Llombart-Cussac, A., Lluch, A., Ribelles, N., Munárriz, B., Martin, M.: Improvement of Breast Cancer Relapse Prediction in High Risk Intervals using Artificial Neural Networks. Breast Cancer Research and Treatment 94, 265–272 (2005)
3. Subirats, J.L., Jerez, J.M., Franco, L.: A New Decomposition Algorithm for the Synthesis and Generalization of Boolean Functions. Submitted (2007)
4. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools with Java implementations. Morgan Kaufmann, San Francisco (2000)
5. Quinlan, J.R.: C4.5: Program for Machine Learning. Morgan Kaufmann, San Mateo (1993)
6. Tilbury, J.B., Van Eetvelt, W.J., Garibaldi, J.M., Curnsw, J.S.H, Ifeachor, E.C.: Receiver operating characteristic analysis for intelligent medical systems-a new approach for finding confidence intervals. IEEE Trans. on Biomedical Engineering 47, 952–963 (2000)
7. Lisboa, P.J.G, Vellido, A., Wong, H.: Outstanding Issues for Clinical Decision Support with Neural Networks. In: Artificial Neural Networks in Medicine and Biology, pp. 63–71. Springer, London (2000)
8. Fawcett, T.: ROC graphs: Notes and practical considerations for researchers, Technical report, HP Laboratories (2004)
9. Saphner, T., Tormey, D.C., Gray, R.: Annual hazard rates of recurrence for breast cancer after primary therapy. Journal of Clinical Oncology 14, 2738–2746 (1996)