A Model For Prognosis of Early Breast Cancer


(1) Group of Research in Images Analysis and Artificial Intelligence
Departamento de Lenguajes y Ciencias de la Computación, University of Málaga
(2) Oncology Service from the Hospital Clínico Universitario of Málaga,
Málaga, 29071, SPAIN

Abstract: - The prediction of clinical outcome of patients after breast cancer surgery plays an important role in medical tasks like diagnosis and treatment planning. These kinds of estimations are currently performed by clinicians using non-numerical techniques. Artificial neural networks are shown to be a powerful tool for analyzing data sets where there are complicated nonlinear interactions between the input data and the information to be predicted. In this paper, we study the approximation of the a posteriori probability of Bayes by feedforward three-layer neural networks. This study is applied for the prediction of patients relapse probability using clinical-pathological data (tumor size, patient age, estrogen receptors, etc.) from the Medical Oncology Service of the Hospital Clínico Universitario of Malaga, Spain. Different network topologies and learning parameters are investigated to obtain the best prediction accuracy. The actual results show as, after training process, the final model is appropriate to make predictions about the relapse probability at different times of follow-up.

Key-Words: - Backpropagation algorithm, Bayes error, diagnostic systems, multi-layer feed-forward neural networks, medical diagnosis.

1 Introduction
Prediction tasks are among the most interesting areas in which to implement intelligent system. More exactly, prediction is an attempt to accurately forecast the evolution or outcome of a specific situation, using as input information obtained from a concrete set of variables that describe that situation.

A problem often faced in clinical medicine is how to reach a conclusion about prognosis of patients when presented with complex clinical and prognostic information. The clinician usually makes decision based on a simple dichotomization of variables into favorable and unfavorable classification [9]. We analyze the decision making process existing when patients with primary breast cancer should receive a certain therapy to remove the primary tumor. In this point is very important the likelihood that the patient will suffer a recurrence of her disease, so that the risks and expected benefits of specific therapies can be compared.

Neural Network are a form of artificial intelligence that have found application in a wide range of problems [6,10,11] that have given in multiple cases superior results to standard statistical models [13]. Baxt [2] demonstrated the predictive reliability of artificial neural networks model in medical diagnosis. In this case, we utilize the ability of neural networks to recognize complex and highly non-linear relationships, such as are likely to characterize medical circumstances.

Some authors [12,7] have modeled systems for outcome prediction after surgery of patients with breast and lung carcinoma. They use neural networks to perform survival analysis together with different survival estimators to deal with censored data of patients. This implies that prognostic factors, for example in early breast cancer with adjuvant therapy after surgery, are not time-dependent, but this is not really true. This is, the strength of prognostic factor is not the same for different time intervals. Different techniques for survival estimation, as Kaplan-Meier analysis [8] and Cox-Regression modeling [3] assumes that the strength of a prognostic factor do not change over time. Also, the existence of a “peak” of recurrence in the distribution of relapse probability [1] demonstrates that the recurrence probability is not the same over time. So that, we proposed a new system approach based on specific topologies of neural networks for different time intervals during the follow-up time of the patients.

This paper is organized as following: in Section 2 we present the experimental material to be used; in Section 3 the decision rule, the prognosis model and the results are proposed; the final conclusions and future works are described in Section 4.
2 Experimental Material

The patient data used in the analysis was obtained from the Medical Oncology Service of the Hospital Clinico Universitario of Malaga, Spain. The records of 875 patients with breast tumors for the ten last years were respectively registered in a database. Data corresponding to each patient are structured in 85 fields containing information about postsurgical measurements, particulars and type of treatment. Medical experts pointed out the importance of a set of prognostic factors selected from all the fields above mentioned. After designing several system with different combinations of these prognostic factors, the set of variables composed by patient age, tumor size, number of axillary lymph nodes, estrogen receptors, and grade of tumor (these two last indicators are qualitatively measured) lead us to achieve the best prediction accuracy: Table 1 shows the mean, range and standard deviation for all the variables representing the prognostic factors.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Range</th>
<th>Mean</th>
<th>STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>24-89</td>
<td>54.59</td>
<td>12.27</td>
</tr>
<tr>
<td>Estrogen Receptors</td>
<td>1,2</td>
<td>1.53</td>
<td>0.499</td>
</tr>
<tr>
<td>Grade</td>
<td>1,2,3</td>
<td>2.214</td>
<td>0.63</td>
</tr>
<tr>
<td># Axillary lymph nodes</td>
<td>0-34</td>
<td>2.52</td>
<td>4.29</td>
</tr>
<tr>
<td>Tumor size</td>
<td>0-230</td>
<td>31.15</td>
<td>21.03</td>
</tr>
</tbody>
</table>

Table 1. Summary of patient data: means, standard deviation and ranges.

3 The Diagnosis System

3.1 Approximating Bayes’ Decision Rule

The problem presented is: given a patient, will he suffer a postsurgical relapse at any period during his follow-up time? We need a decision rule to solve it. When minimizing the correct classification probability is desired, the best decision rule has been demonstrated to be Bayes’ rule [4]

\[
\phi_i(x) = \begin{cases} 1 & \text{if } p(C_i / x) \geq p(C_k / x) \forall k \\ 0 & \text{in otherwise} \end{cases}
\]

where \( \phi_i \) is the probability of classify the pattern \( x \) in class \( C_i \). So, we are going to determine the correct classification probability of Bayes in our problem. Let \( p(C_i) \) be the a priori probability of class \( C_i \) where \( i = 1 \) identify the class “relapse” and \( i = 2 \) the class “non-relapse” and let \( p_{ii} \) be the conditional probability of classify a pattern of \( C_i \) in \( C_i \), then the correct classification probability is given by

\[
p = p(C_1) \cdot \int_{\mathbb{R}^n} \phi_1(x) \cdot p(x / C_1) \, dx + p(C_2) \cdot \int_{\mathbb{R}^n} \phi_2(x) \cdot p(x / C_2) \, dx
\]

\[
= \int_{x \in \mathbb{R}^n} p(C_1) \cdot p(x / C_1) \, dx + \int_{x \in \mathbb{R}^n} p(C_2) \cdot p(x / C_2) \, dx
\]

\[
= p(C_1) + \int \frac{[p(C_i / x) - p(C_i / x)]}{\pi} \cdot p(x) \, dx
\]

\[
= p(C_1) + \int \frac{[1 - 2 p(C_i / x)]}{\pi} \cdot p(x) \, dx \quad (1)
\]

In an analogous form we have

\[
p = p(C_2) + \int \frac{[2 p(C_i / x) - 1]}{\pi} \cdot p(x) \, dx \quad (2)
\]

From Eq (1) and Eq (2) we obtain

\[
p = \frac{1}{2} + \int \frac{p(C_i / x) - \frac{1}{2}}{\pi} \cdot p(x) \, dx \quad (3)
\]

Therefore \( p \geq \max \{p(C_1), p(C_2)\} \) that gives an idea about the lower value for the correct classification probability.

The a posteriori probability density function of \( p(C_i / x) \) is unknown in the presented problem, so that, we should estimate it in order to obtain an approximated correct classification probability. Funahashi [5] shown that in a three-layer neural network, using in a back-propagation algorithm as teaching signals 1 when the input pattern belongs to class \( C_1 \) and 0 when belongs to class \( C_2 \) and if the learning proceeds ideally, the output of the network tends to the a posteriori probability density function \( p(C_i / x) \). That is when the mean-squared error decreases to its infimum, we have

\[
p(C_i / x) = F(x, t, w) \quad (4)
\]

where \( F(x, t, w) \) is the network output for an input pattern \( x \) and \( t \) and \( w \) are the synaptic weight matrices.

Hence, the estimated correct classification Bayes’ probability is given by the expression

\[
p = p(C_1) \cdot p_{11} + p(C_2) \cdot p_{22}
\]
\[ \phi(x) = \begin{cases} 1 & \text{if } F(x, t, w) \geq 1/2 \\ 0 & \text{if } F(x, t, w) < 1/2 \end{cases} \]  

(5)

which gives the probability of classifying the pattern \( x \) in class \( C_i \). If \( \phi(x) = 1 \), pattern \( x \) is classified in class \( C_1 \) and if \( \phi(x) = 0 \) is classified in \( C_2 \).

From Eq. (3) and Eq. (4), we obtain that the estimated probability for correct classification is given by the expression

\[ \hat{p} = \frac{1}{2} + \frac{1}{n} \sum_{i=1}^{n} F(x_i, t, w) - \frac{1}{2} \]  

(6)

where \( n \) is the number of patients. The correct classification Bayes’ probability estimated in Eq. (6) is the superior bound we can reach in this problem, that is, it gives us an idea of the classification difficulty in where the best decision rule could give at most that probability.

3.2 The proposed model

Prognostic factors in early breast cancer with adjuvant therapy after surgery are time-dependent. This is, the strength of prognostic factor is not the same for the first 10 months as, for example, the 50-60 months interval. Different techniques for survival estimation, as Kaplan-Meier analysis [8] and Cox-Regression modeling [3] assumes that the strength of a prognostic factor do not change over time, and this is not the case in the “real world”. On the other hand, the recurrence probability is not the same over time, since the existence of a “peak” of recurrence in the distribution of relapse probability has been demonstrated empirically [1].

Considering this and the justification of the proposed decision rule in Eq. (5), a solution scheme is proposed based on specific topologies of neural networks for different time intervals during the follow-up time of the patients (Table 2). It consists of a neural networks system and a threshold unit to implement the decision making process (see Fig. 1). The neural networks system computes an attributes set from a patient record and gives a value corresponding to the a posteriori probability of relapse for that patient. The threshold unit then outputs a class according to the proposed decision rule in Eq. (5).

![Figure 1: Proposed diagnosis system.](image)

All networks have three layers (input, hidden and output) and use the hyperbolic tangent function in the hidden layer and the logistic function in the output layer.

A crucial aspect of doing learning and prediction analysis with a neural network system is to split the data base into two independent sets which will be used to train the neural network and to validate its predictive performance. During training the data vectors of the training set are repetitively presented to the network, that attempts to generate a 1 at the output unit when the survival status of the patient is relapse, and a 0 when the status is non-relapse. Connection weights are changed using a Levenberg-Marquardt errors back propagation algorithm.

Giving the information to the neural network input layer requires an information preprocessing process. First, it is important to normalize all the prognostic factors ranges to lie within the central range of the hidden layer transfer function in the neural network (-1.0 and 1.0 for the hyperbolic tangent transfer function). Second, studying of the range and distribution of each prognostic variable to remove all the missed values, and to lessen the
impact of outlying points at the extremes of the distribution.

Data subsets corresponding to each time interval studied were selected from the original 875 patients from the Oncology Service database in order to classify them in classes $C_1$ and $C_2$. Given a time interval $I_i$ (see Table 2), patients from the original data set who fall into the two separated classes for this studied interval are selected according to the following rules

1. Patients from Interval $I_i$: Those whose survival status is relapse are selected for class $C_1$. The rest are ignored.

2. Patients from Interval $I_j$ ($j<i$): Those whose survival status was relapse are selected for class $C_2$. The rest are ignored.

3. Patients from interval $I_k$ ($k>i$): All the patients are selected for class $C_2$.

Table 2 shows the final number of individuals selected and the a priori probability of relapse for each time interval studied after applying the above rules.

<table>
<thead>
<tr>
<th>Time Intervals (# months)</th>
<th># of Patients</th>
<th>A priori prob. of relapse</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_1$ (0 – 10)</td>
<td>749</td>
<td>8.2 %</td>
</tr>
<tr>
<td>$I_2$ (10 – 20)</td>
<td>732</td>
<td>12.45 %</td>
</tr>
<tr>
<td>$I_3$ (20 – 30)</td>
<td>654</td>
<td>10.9 %</td>
</tr>
<tr>
<td>$I_4$ (30 – 40)</td>
<td>430</td>
<td>12.25 %</td>
</tr>
<tr>
<td>$I_5$ (40 – 50)</td>
<td>573</td>
<td>11.38 %</td>
</tr>
<tr>
<td>$I_6$ (50 – 60)</td>
<td>498</td>
<td>9.4 %</td>
</tr>
<tr>
<td>$I_7$ (&gt; 60)</td>
<td>498</td>
<td>8.12 %</td>
</tr>
</tbody>
</table>

Table 2. Number of patients selected and the a priori probability of relapse for each time interval.

Weights initialization is crucial in the learning process with artificial neural networks. In order to improve the quadratic error by the end of the learning process, several weights initialization were carried out to get the best set of final values for synaptic weights. The estimation of the model accuracy is obtained from the ten-fold cross-validation method to obtain mean values for probabilities of correct classification probabilities.

3.3 Results

To evaluate the correct classification probability of the proposed system, the estimated probability of Bayes, the a priori probability of the data set, and that obtained by application of our decision rule in Eq. (5) have been simultaneously plotted with the best topologies of neural network found for each time interval (Fig. 2).

Several results are observed in Fig. 2: First, the proposed system (NNCP) always improves the a priori probability (PCP). Here, it is important to point out the difficult of this, given so high values of PCP for each time interval. Besides, this improvement is greater in the most critical interval during the follow-up time of the patients (interval $I_2$). Second, the NNCP is always smaller than BCP and it follows the BCP shape, as was expected. Also, we can observe that the difference between both is not significant, what means that the proposed rule in Eq. (5) is a good estimator of the Bayes’ decision rule.

4 Conclusions and future work

Different topologies of feed-forward neural networks were used to obtain the best prediction accuracy for correct classification probability of patients relapse after breast cancer surgery using clinical-pathological
data. The final prognosis system, based on the decision rule proposed in this work, is to make predictions about the relapse probability at different times of follow-up with a very little error ratio. On the other hand, we advance a Neural Network-based method to estimate the correct classification probability of Bayes, based on the approximation of the a posteriori probability from a distribution using a multi-layer perceptron.

The following stage in this work will be an attempt to find the relative weightings of the different variables that are used to describe the different cases by introducing a methodology based on Genetic Algorithms for the automatic induction of Neural Networks topologies to identify those variables becoming the best prognostic factors to make prognosis at different time intervals.

In a posterior work, we will show the results obtained by integrating artificial neural networks and decision trees based learning rules in the design of the prediction system.

Acknowledgments
We would like to thank the personal from the Oncology Service of the Hospital Clínico Universitario de Malaga for their comments and collaboration in this work.

References: