Neural Network Prediction of Relapse in Breast Cancer Patients

L. Tarassenko¹, R. Whitehouse², G. Gasparini³ and A. L. Harris¹

¹ Department of Engineering Science, University of Oxford, Oxford; ² ICRF Clinical Oncology Unit, Churchill Hospital, Oxford, UK; and ³ Department of Oncology, St. Bortolo Regional Medical Centre, Vicenza, Italy

When a woman diagnosed as having breast cancer has a tumour removed, it is important to try and predict whether she is likely to relapse within, say, the next three years. In this paper, the performance of a neural network classifier trained on a number of prognostic indicators is shown to be better than that of the clinical experts working with the same information. To obtain meaningful statistics with the relatively small dataset available, the network is trained using a modified form of the leave-one-out method. A procedure is also introduced for investigating how much independent information each input parameter contributes. This shows that, in this type of retrospective study, the type of therapy given to the woman does not significantly affect the network's prediction of whether or not she will relapse within three years. Finally, since this problem, in common with many other medical problems, is plagued by a shortage of data, the final section of the paper reports on an investigation of whether or not multi-centre databases might be feasible.

Keywords: Breast cancer; Prediction of outcome; Prognostic indicators; Neural network classifier, Multilayer perceptron; Leave-one-out method

1. Introduction

When a tumour is removed after a woman has been diagnosed as having breast cancer, clinicians will estimate the risk of recurrence to determine which treatment might be most appropriate. The type of tumour, its size and the number of nodes, together with the age of the patient, all represent important data which the clinician uses to assess the patient's prognosis. Depending on this assessment, the woman may be given no specific therapy (if the prognosis is good), hormone therapy, chemotherapy or both (when the prognosis is very poor). Recently, information of a biochemical nature has also become available, for example, the oestrogen receptor (er) and the epidermal growth factor receptor (egfr) status at the time of the tumour removal, but perhaps this makes it even more difficult to reach a conclusion. It has been argued that the clinician usually makes the decision using first a simple dichotomisation of variables into favourable and unfavourable classifications, and then subjectively weighting the factors to reach an overall assessment [1].

The prediction of whether or not a woman will relapse within a given time of the primary tumour being removed is a very hard problem; under the conditions described in this paper, the accuracy of trained experts is no better than 65%-70% (see Sect. 5.1). The aim of this paper is to present a principled approach to the use of a neural network classifier to decide whether or not a woman will relapse within three years. We have used three years rather than five years because it is only since 1989 that the Imperial Cancer Research Fund (ICRF) Unit in Oxford began to collect all of the relevant data; even then, as will be seen in the next section, there are many patients with missing data. We will first use visualisation techniques to obtain an idea of the complexity of the task, but the main objective of the paper is to assess the suitability of a multilayer perceptron (MLP) for predicting whether or not a woman will relapse. The MLP
will be trained using a modified version of the leave-one-out approach, and we will present a strategy for the estimation of the importance of input variables, principally the type of therapy given to the patient. Finally, we will consider whether or not it is worthwhile, or even possible, to combine data from two different centres.

2. Selection of Database

As of June 1995, the ICRF Unit at the Churchill Hospital in Oxford had assembled a database of 658 patients for whom it had 3-year follow-up data. Of these, 177 were known to have relapsed within three years of the primary tumour being removed. However, not all of the parameters were recorded in the early days (this being especially true of the er status and the egfr status), and so there are only 325 patients with no missing data in the database, including 100 cases of relapse within three years. The decision was made early on in this study to concentrate solely on these patients for whom all the parameters had been recorded.

There are three main types of primary tumours: ductal, lobular and 'others'. The overwhelming majority of patients have a ductal tumour (81% of the non-relapse cases and 92% of the relapse cases), and so it was decided to concentrate on this type of cancer for two reasons: firstly, with a database of few examples at this stage, it is important to keep the number of inputs as low as possible to minimise the number of free parameters in the neural network model; coding the type of tumour would have necessitated three extra inputs. Secondly, it is not known whether the correlations between the prognostic indicators are likely to be different for the different types of tumours, and hence it is advisable at first to concentrate on the dominant type of cancer on its own.

The distribution of patients with a ductal tumour is shown in Table 1. There are 183 non-relapse cases (out of the 225 non-relapse cases in the database) and 92 relapse cases (out of 100). Ductal carcinoma is normally graded into three categories of differentiation, grade I being most similar to normal breast in structure and grade III having no detectable tubules or differentiation. For the purposes of this study, grades I and II were grouped together and categorised as being well differentiated (ductalO for ordinary differentiation) and grade III was categorised as being poorly differentiated (ductalP). The number of women with ductalP tumours is evenly distributed between the two cases, but there are far more women with a ductalO tumour who do not relapse. Since it is not known at this stage whether the frequency of occurrence in the database is in any way indicative of the prior probabilities (amongst other things, there is the complication introduced by rejecting all patients with one or more missing parameters), an approximately balanced data set was constructed instead; 50 women with a ductalO tumour who did not relapse were chosen at random from the 133 possible cases to give a database of 192 women with ductal tumours (100 with no relapse and 92 who do relapse within three years). This data set will be the focus of all the investigations reported in this paper, until new data are considered in the final section.

3. Preliminary Analysis

It was argued in the introduction that when a clinician decides which form of therapy, if any, should be given to a woman who has just had a breast cancer tumour removed, the decision is made, at least partly, on the basis of the five parameters which have already been listed: age of the patient, tumour size, number of nodes, er status and egfr status. The mean values and standard deviations of these prognostic indicators for the data set of 192 cases are given in Table 2 below. Note that the er and egfr parameters are the natural logarithms of the original values; the tumour size in given in centimetres.

It is clear from the values of the standard deviations with respect to the means, that the distributions are far from Gaussian, and hence that the classical quadratic or linear discriminant methods (which assume multivariate normal distributions) will not perform well on this data set.

3.1. Sammon Mapping

The difficulty of the classification task (predicting whether or not a woman will relapse on the basis