Choong, P.L., deSilva, C.J.S. and Attikiouzel, Y. (1997) Predicting local and distant metastasis for breast cancer patients using the Bayesian neural network. In: Proceedings of the 13th International Conference on Digital Signal Processing , 1997. DSP 97., 1997 Digital Signal Processing Proceedings, DSP 97, 2 - 4 July, Santorini, Greece, pp. 83-88.

Predicting Local and Distant Metastasis for Breast Cancer Patients using the Bayesian Neural Network

Poh Lian Choong Christopher J.S. deSilva Yianni Attikiouzel
Centre for Intelligent Information Processing Systems
Department of Electrical and Electronic Engineering
The University of Western Australia
Nedlands WA 6907, Australia
pohlian@ee.uwa.edu.au chris@ee.uwa.edu.au yianni@ee.uwa.edu.au

Abstract

This paper presents a predictive accuracy comparison between the Multivariate Logistic Regression (MLR) and the Bayesian Neural Network (BNN). The latter is presented in this paper as an alternative to the MLR (MLR). The MLR and BNN have been used to identify early breast cancer patients with high risk of tumour recurrence at the time of initial resection.

1 Introduction

Several issues have been raised by statisticians and medical personnel in relation to the application of the standard Multilayer Perceptron to the analyss of medical data. The problem lies in the application of the Artificial Neural Networks (ANN) paradigm, issues such as the which input variable is important and the correct number of layers and the number of units in each layer in order to obtain a certain level of performance.

To resolve the problem of selecting the optimum design choices for the MLP, we applied Bayes' Theorem, which embodies the philosophy of William Occam, to provide a framework for selecting the optimum ANN architecture. The architecture is optimal in the sense that preference is given to the simplest model (least number of layers and units) which adequately models the training data. Model comparison can be done by evaluating a quantitative value termed the evidence. The evidence calculated for each ANN model takes into consideration the goodness of fit and the complexity of the model. The principle penalises complex models and favours the simplest model. Using the Bayesian Neural Network (BNN), we were able to address some of the drawbacks of the standard MLP.

Bayesian methods for inductive inference were developed in detail by a Cambridge geophysicist, Sir Harold Jefferys [5]. Bayes' theorem is regarded as a form of common sense reasoning, providing the framework to manipulate probability distributions. But to apply Bayesian reasoning, firstly we need to transform the medical information into a numerical probability distribution using some other principle. This can be achieved using the standard statistical models such as the Multivariate Logistic Regression (MLR) or ANNs. Bayes' framework can also be used to determine the significance of each individual risk factor to the outcome. The framework, termed *Automatic Relevance Determination* (ARD), is due to published works of MacKay in 1991 - 1992 [6, 7, 8], Gull [3, 4].

About 7000 Australian womens are diagnosed with early breast cancer annually [2]. In many of these patients, breast cancer is a systemic disease at diagnosis and is therefore not curable by surgical removal of the primary tumour alone. Breast cancer is a heterogeneous disease, resulting in a wide range of treatment options. These treatments vary widely in toxicity, from relatively harmless (such as tamoxifen) to highly aggressive experimental therapy (such as bone narrow transplantation). Decisions about which patients to treat with these different forms of adjuvant therapies require that we confront two important issues. Firstly, what is an acceptable risk of recurrence, that is, a risk so low as to argue against the need for systemic therapy. Accurate assessment of the probability of recurrence is therefore essential in deciding the appropri-

[0]This research was supported by the Cancer foundation of Western Australia.

0-7803-4137-6/97/$10.00©1977 IEEE

ate adjuvant treatment for individual patients [1]. Secondly, the question arises as to whether we have the prognostic factors to predict the risk of recurrence with a high degree of accuracy.

This paper focuses on the question of how we can best identify node positive patients who will have a high or low probability of recurrence at the early stage of the disease. We investigated the use of four risk factors previously analysed using statistical methods by Seshadri and associates [9] to predict the risk of recurrence for individual patients following the removal of the primary tumour and initial adjuvant therapy. Risk prediction allows identification of patients to be considered for additional therapy or to select appropriate treatment. Patients expected to have low risk of recurrence will be spared from additional or delayed toxic effects

2 The Bayesian Neural Network Formalism

David MacKay [6, 7, 8] has provided a comprehensive and detailed description and analysis of the incorporation of the Bayesian inference and evidence framework with the MLP network. The BNN has been implemented by MacKay [6] using a deterministic method involving Gaussian approximation.

The formulation of the BNN involves solving an iterative top-down approach for four levels of inference as shown in Figure 1 [6].

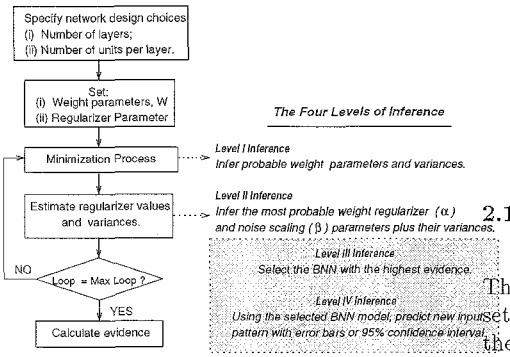

Figure 1: Four levels of BNN inference

The ultimate goal in the modelling process is to be able to predict the outcome for new patients based on the information contained in a set of risk factors. The accuracy of the prediction will be dependent on the model constructed. Therefore, only the simplest model that best describes the training data should be used for Level IV inference. Level III of the inference process involves comparing and ranking preferences for alternative ANN models using the evidence framework, $P(\mathcal{D} \mid \mathcal{M}_i)$. The evidence from each trained ANN model can only be determined if we have inferred the most probable weight (ω_{MP}), regulariser (α) and noise scale parameters (β) determined from inference level I and II. The following sections will describe each level of inference in detail.

2.1 Bayesian Inference of the weight parameters (Level I)

Bayesian analysis involves using the output of the network, $\widehat{\mathbf{g}}(\mathbf{x})$ to construct the likelihood function, $P(\mathbf{y} \mid \mathbf{x}, \omega)$ for the training data $\mathcal{D} = (\mathbf{x}, \mathbf{y})$. We want to train the BNN network to give us the most probable weight parameter, ω_{MP}, given the data $(\mathbf{x}, \mathbf{y})$, network configuration $(\mathcal{M}_i)$ and some scaling parameters which will be discussed later in this chapter. Using Bayes' rule, the posterior probability of the weight parameter is;

$$P(\mathcal{D} \mid \omega, \mathcal{M}_i)\ P(\omega \mid \mathcal{D}, \mathcal{M}_i) = P(\omega \mid \mathcal{D}, \mathcal{M}_i)\ P(\mathcal{D} \mid \mathcal{M}_i) \quad (1)$$

$$P(\omega \mid \mathcal{D}, \mathcal{M}_i) = \frac{P(\mathcal{D} \mid \omega, \mathcal{M}_i)\ P(\omega \mid \mathcal{D}, \mathcal{M}_i)}{P(\mathcal{D} \mid \mathcal{M}_i)} \quad (2)$$

where $P(\omega \mid \mathcal{D}, \mathcal{M}_i)$ is the posterior probability of the ANN weight parameters, ω.

Infering the most probable weights involves:

1. Computing the likelihood, $P(\mathbf{y} \mid \mathbf{x}, \omega, \mathcal{M}_i)$ of the data.

2. Computing the prior probability of the weight parameters, $P(\omega \mid \mathcal{D}, \mathcal{M}_i)$

2.1.1 Computing the likelihood, $P(\mathcal{D} \mid \omega, \mathcal{M}_i)$

The ANN training procedure is used to find a set of weights that maximises the likelihood of the training data, $\mathcal{D}$.

$$\overset{Max}{\omega \in \mathcal{W}}\ P(\mathcal{D} \mid \omega, \mathcal{M}_i) = \overset{Max}{\omega \in \mathcal{W}} \prod_{j=1}^{M} P(D_j \mid \omega, \mathcal{M}_i) \quad (3)$$

The MLP network is usually trained to best fit the training data, $\mathcal{D}$ by minimising a quadratic

error function or the Euclidean norm,

$$E_D = \frac{1}{2}\sum_{j=1}^{M}[y_j - \widehat{g}(x_j, \omega, \mathcal{M}_i)]^2 \quad (4)$$

The functional form of the likelihood model can be reduced to a Gaussian distribution.

$$P(\mathcal{D} \mid \omega, \beta, \mathcal{M}_i) = \frac{1}{(2\pi\sigma^2)^{\frac{M}{2}}} exp\left[-\frac{E_D}{\sigma^2}\right] \quad (5)$$

where $\beta = \frac{1}{\sigma^2}$ is the noise level and $Z_D(\beta) = \frac{2\pi}{\beta}^{M/2}$.

MacKay [6, 7] proposed the use of a "weight-decay" or *regularising term* (α) to prevent weights of irrelevant units from growing too large. The regulariser term forms the smoothing parameter of the ANN model. Inclusion of the regulariser term penalises large weights in the training process.

Training the ANN network to maximise the likelihood of the training data can also be viewed as inferring the most probable weight parameters;

$$P(\omega \mid \mathcal{D}, \beta, \alpha, \mathcal{M}_i) = \frac{P(\mathcal{D} \mid \omega, \beta, \mathcal{M}) P(\omega \mid \alpha, \mathcal{M}_i)}{P(\mathcal{D} \mid \alpha, \beta, \mathcal{M}_i)} \quad (6)$$

where $P(\mathcal{D} \mid \alpha, \beta, \mathcal{M}_i)$ is the normalisation constant and α is the regularising term.

2.1.2 Prior probability of the weights

The only consistent prior for the weight parameter, ω is of the Gaussian form:

$$P(\omega \mid \alpha, \mathcal{M}_i) = \frac{exp(-\alpha E_\omega)}{Z_\omega(\alpha)} \quad (7)$$

where $Z_\omega(\alpha) = \int exp(-\alpha E_W) d^k\omega$ and $E_W = \frac{1}{2}\sum \omega^2$.

Substituting equations (5) and (7) into equation (6) gives a posterior weight distribution:

$$P(\omega \mid \mathbf{y}, \mathbf{x}, \beta, \alpha, \mathcal{M}_i) = \frac{exp(-[\beta E_D + \alpha E_W])}{Z_\Phi(\alpha, \beta)} \quad (8)$$
$$= \frac{exp(-\Phi(\omega))}{Z_\Phi(\alpha, \beta)}$$

where $Z_\Phi(\alpha, \beta) = \int exp(-[\beta E_D + \alpha E_W]) d^k\omega$ and training the ANN by minimising the objective function, $\Phi(\omega) = \beta E_D + \alpha E_W$, would infer the most probable weight, ω_{MP}.

2.2 Selecting the optimal regulariser value, α (Level II)

Solving the objective function, $\Phi(\omega)$ involves determining the values of the regulariser term α and the noise level β. The optimal α and β term can be found by differentiating, $\log P(\mathbf{y} \mid \mathbf{x}, \alpha, \beta, \mathcal{M}_i)$ with respect to α and β and setting these equations to zero.

$$\log P(\mathbf{y} \mid \mathbf{x}, \alpha, \beta, \mathcal{M}_i) \quad (9)$$
$$= -\Phi_{MP} + \frac{k}{2}\log(2\pi) - \frac{1}{2}\log(det\mathcal{A}) - \log Z_\omega(\alpha) - \log Z_D(\beta)$$

2.3 Evaluating the evidence (Level III)

Level III inference involves selecting the ANN model with the highest recorded evidence, $P(\mathcal{D} \mid \mathcal{M}_i)$. The posterior distribution encapsulates all the information about $\mathcal{M}_i$ given the information of the training data, $\mathcal{D}$. Therefore, the evidence for the model, $\mathcal{M}_i$, is:

$$P(\mathbf{y} \mid \mathbf{x}, \mathcal{M}_i) = \int P(\mathbf{y} \mid \mathbf{x}, \alpha, \beta, \mathcal{M}_i) P(\alpha, \beta \mid \mathcal{M}_i) d\alpha d\beta \quad (10)$$

3 Breast Cancer Prognosis

The data set used in this investigation relates to 351 women in South Australia and Western Australia diagnosed with breast carcinoma between 1987 and 1992. For all patients, their diagnosis was confirmed by biopsy and treated either by *total* (75% of patients) or *partial* (25% of patients) mastectomy. All patients had had axillary lymph node clearance with positive confirmation of metastases therein. The four criteria for inclusion in this study are:

1. Only node positive patients are considered.
2. Patients with stage-IV disease or for whom axillary clearance was not performed were excluded.
3. Patients without recurrence are required to have a minimum follow-up period of 18 months.
4. Information on all the risk factors considered is available.

This study concentrates only on the significant predictive property of risk factors to predict local or distant recurrence for the node positive

patient within eighteen months after diagnosis. A consecutive series of 351 node-positive patients with complete axillary dissection were available for this case study. Detailed descriptions regarding the preparation of breast sample for hormone receptor and Cathepsin-D analysis have been given in [9].

3.1 Risk Factors

Histopathological features considered in this case study are similar to the study conducted by Seshadri [9].In this study, apart from analysing the prognostic values of the respective risk factors, consideration was also given to predicting the risk of relapse for individual patients. The risk factors used in the study are Tumour Size (TS), Number of Nodes, Estrogen Receptor (ER) and Cathepsin-D (Cath-D).

The breast carcinoma data were analysed using both the Bayesian Neural Network (BNN) and the Multivariate Logistic Regression (MLR) methods. Patients were assigned randomly and independently from the data set to a training set used to construct the model and a testing set used to evaluate the performance of the model. A random sample of 184 patients (150 MET- and 34 MET+) was selected to provide the training data, leaving 167 patients (137 MET- and 30 MET+) as a test set on which to validate the predictions.

4 Results

4.1 Construction of the Standard MLR, Bayesian MLR and BNN model

For the construction of the Standard MLR models, the standard multivariate logistic regression method was applied. The Bayesian MLR model was constructed using the BNN network with no hidden nodes and one output node with sigmoidal activation function. This BNN network was trained to maximise the likelihood of the training data. This is similar in principle to the multivariate logistic function, except in this case, regulariser terms are attached to each input. The parametric function of this BNN model is equivalent to the standard MLR model.

In the case of the BNN model, one hidden layer was used in the modelling process. The sigmoidal activation function was applied to all the units with the exception of the input units. We have noted in the earlier sections that the BNN network is an integration of the Multilayer Perceptron and Bayes Theorem. Using the BNN network, we can use the Occam framework to determine the evidence for each ANN network. The evidence framework allows us to rank preferences for alternative BNN models. Several BNN networks with different numbers of hidden units were trained and the log evidence, $P(D \mid M_i)$, of each trained BNN model was evaluated. The evidence for each BNN model trained is shown in Figure 2. Since the training process of the BNN network involves initialising

Since the training process of the BNN network involves initialising the weight parameters and the regulariser terms to small random values, we decided to retrain the BNN network three times for each design choice to the same tolerance level and evaluate the evidence each time. The results tabulated in Figure 2 show a rapid increase of $P(\mathbf{y} \mid \mathbf{x}, \mathcal{M}_i)$ with increasing numbers of hidden units. Additional hidden units improve the fit but also increases the complexity of the model. The BNN incorporating the Occam Razor selected the optimum BNN model to best fit the data with only 9 hidden units. Therefore the BNN architecture used for the classification analyses will have the following architecture, 4 input units, 1 hidden layer with 9 units and 1 output unit.

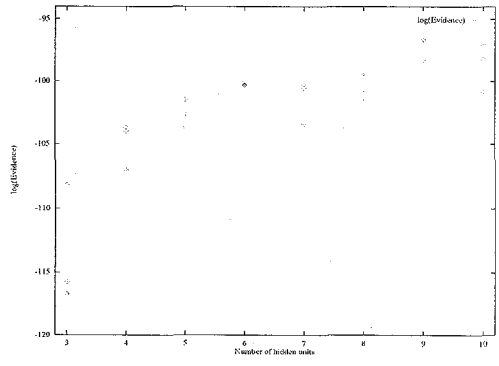

Figure 2: Relationship between the number of nodes and log of the evidence, $P(\mathbf{y} \mid \mathbf{x}, \mathcal{M}_i)$

In the multivariate analysis, the results from both the MLR and BNN model tabulated in Table 1 show that only the pathologic status of the lymph nodes and the ER concentration were statistically significant in the multivariate model. In the MLR model, both these risk factors have $P < 0.05$ and in the BNN model, the regulariser term or the decay term associated with each in-

Risk Factors	Std MLR (P)	Bayes (α)
1. Tumour Sze	0.154	0.358
2. No of Nodes Involved	0.020	0.029
3. Estrogen Receptor	0.007	0.016
4. Cathepsin-D	0.884	0.310

Table 1: Significance of Risk Factors

Testing Data	Predicted					
	Std MLR		Bayes MLR		BNN	
Actual	M-	M+	M-	M+	M-	M+
M-	79	58	80	57	84	53
M+	10	20	8	22	9	21
Total	89	78	88	79	93	74

Table 2: Comparison of the classification accuracy.

put is small in comparison to the α values for tumour size and Cathepsin-D (larger by a factor of 10). Note that in both the MLR and BNN analyses, risk factors which are uncorrelated to the outcome will be inferred large P and α values.

4.2 Forced Classification Accuracy

Patients were stratified into two risk groups, M+ and M-. Table 2 tabulates the predicted status of individual patients. The BNN has higher predictive accuracy for both M+ and M- patients in comparison to the MLR and the Bayesian MLR model.

5 Discussion

The analyses carried out in this case study concentrates on two things. Firstly, can we predict the probability of recurrence for the node-positive patient on the basis of information about tumour characteristics. Secondly, to compare the predictive accuracy of the MLR and BNN network since the use of the BNN network is a novel approach in this area. The MLR, Bayesian MLR and BNN models were constructed using 184 patients (training data) and the remaining 167 patients were used to assess their predictive quality.

The BNN embodies the Occam principle and Bayes Theorem to provide a quantitative assessment for ranking alternative ANN models. Using the evidence framework, a BNN network with 9 hidden units was selected. The BNN network with 9 hidden units achieved the highest evidence as shown in Figure 2. We can now do away with the ad-hoc method of selecting MLP models. Apart from providing the evidence framework for ranking ANN models, we can also determine which risk factors are correlated to the prediction of recurrence. This knowledge has previously been embedded in the distributed weights of the MLP model. Using a regulariser term (α) for each input, the α associated with irrelevant inputs will be given large values to prevent these inputs from affecting the resulting prediction. In the multinomial model, only two variables were found to be statistically significant independent predictors. They are number of nodes involved ($\alpha = 0.029$) and the ER concentration ($\alpha = 0.016$). The same two variables were also found to be statistically significant in the MLR model, $P = 0.02$ and $P = 0.007$, respectively. Both the BNN and the MLR models found the risk factors tumour size and Cathepsin-D to be uncorrelated to prediction of recurrence. Note that the prognostic significance of ER may be due to its relationship with response to tamoxifen administered to some patients.

he MLR and the BNN models were also tested on 167 patients not used to construct the model to assess their predictive quality. The BNN network predicted more accurately for both MET+ patients (70.0%) and the MET- patients (61.3%). In comparison, the MLR prediction accuracy was 57.6% and 66.7% for MET- and MET+ respectively. The higher percentage of patients correctly classified as having MET+ indicates that the adjuvant therapy selected for individual patients only prolongs the disease free recurrence for 38.7% of the MET- patients. The analysis carried out considered only four risk factors to predict the probability of relapse and was restricted to patients who had had some form of adjuvant therapy.

In view of the small sample size, further verification of the BNN model is required before it can be applied to assess the probability of recurrence in larger populations.

6 Conclusion

The novel analysis using the Bayesian Neural Network has been shown to eliminate some of the drawbacks of the standard MLP network. The evidence framework and the automatic rel-

evance determination provide ways to determine the optimum network size plus identifying inputs which are independent predictors of the outcome. The BNN model was more accurate (higher likelihood) and achieved higher predictive value in comparison to the multivariate logistic regression models. This case study concentrated on assessing the risk of recurrence for node positive patients for whom some form of adjuvant therapy has been given. The same modelling procedure can be used to assess the risk of recurrence for node negative patients or for node positive patients for whom adjuvant therapy have not been administered. This modelling process will enable clinicians to assess the risk of recurrence given the information about tumour characteristics and to select appropriate treatment for both high risk node negative patients and node positive patients.

References

[1] N.E. Davidson, and M.D. Abeloff, "Adjuvant Therapy of Breast Cancer," *World J. Surg.*, **18** , pp. 112 - 116, 1994.

[2] P. Fitzgerald, N. Thomson and J. Thompson, "Cancer incidence and mortality in Western Australia 1990," Health Services Statistics & Epidemiology Branch, Health Department of Western Australia, 1993.

[3] S.F. Gull, "Bayesian Inductive Inference and Maximum Entropy," *Maximum Entropy and Bayesian Spectral Analysis and Estimation Problems*, C. R. Smith and G.J. Erickson (eds.), D. Reidel Pub. Comp., pp. 53 - 74, 1989.

[4] S.F. Gull, " Development in Maximum Entropy Data Analysis," *Maximum Entropy and Bayesian Methods, Cambridge, 1988*, Kluwer, pp. 53 - 71, 1989.

[5] H. Jefferys, *Theory of Probability*, Second Edition, Oxford, Clarendon Press, 1948.

[6] D.J.C. MacKay, *Bayesian Methods for Adaptive Models*, **PhD Thesis** , California Institute Tech, 1992.

[7] D.J.C. MacKay, "Bayesian Non-linear Modeling for the Prediction Competition," submitted to MaxEnt (1993).

[8] D.J.C. MacKay, "Hyperparameters: optimize, or integrate out?," submitted to Neural Computation (1993).

[9] R.Sheshadri, D.J. Horsfall, F. Firgaira, K. McCaul, V. Setlur, A.H. Chalmers, R. Yeo, D. Ingram, H. Dawkins, R. Hahnel and the South Australian Breast Cancer Study Group, "The Relative Prognostic Significance of Total Cathepsin D and Her-2/neu Oncogene Amplification in Breast Cancer," *Int J Cancer*, **56**, pp. 61 - 65, 1994.