PROGNOSTIC COMPARISON OF STATISTICAL, NEURAL AND FUZZY METHODS OF ANALYSIS OF BREAST CANCER IMAGE CYTOMETRIC DATA

H. Seker[1], M. Odetayo[1], D. Petrovic[1], R.N.G. Naguib[1],
C. Bartoli[2], L. Alasio[2], M.S. Lakshmi[3] and G.V. Sherbet[4]
[1]BIOCORE, School of Mathematical and Information Sciences, Coventry University, Coventry, UK
[2]Istituto Nazionale per lo Studio e La Cura dei Tumori, Milan, Italy
[3]Cancer Research Unit, The Medical School, University of Newcastle upon Tyne, UK
[4]The Institute for Molecular Medicine (IMM), Huntington Beach, CA, USA

Abstract - **This paper aims to predict a breast cancer patient's prognosis and to determine the most important prognostic factors by means of logistic regression (LR) as a conventional statistical method, multilayer backpropagation neural network (MLBPNN) as a neural network method, fuzzy *K*-nearest neighbour algorithm (FK-NN) as a fuzzy logic method, a fuzzy measurement based on the FK-NN and the leave-one-out error method. The data used for breast cancer prognostic prediction were collected from 100 women who were clinically diagnosed with breast disease in the form of carcinoma or benign conditions. The data set consists of 7 image cytometric prognostic factors and 2 corresponding outputs to be predicted: whether the patient is alive or dead within 5 years of diagnosis. The LR stratified a 5-factor subset with a prognostic predictive accuracy of 82%, while the highest predictive accuracy of the MLBPNN was 87% obtained from two subsets. In this study, the FK-NN yielded the highest predictive accuracy of 88% achieved by eight different subsets, of which the subset with the highest fuzzy measurement was {tumour histology, DNA ploidy, SPF, G_0G_1/G_2M ratio}. Although the three methods resulted in different models, the results suggest that tumour histology, DNA ploidy and SPF, which are included in all three methods, may be the most significant factors for achieving accurate and reliable breast cancer prognostic prediction.**

Keywords - **Oncology, fine-needle aspirates, survival analysis, knowledge based systems, logistic regression, artificial neural networks, fuzzy *K*-Nearest neighbour classifier.**

I. INTRODUCTION

Prognosis of cancers is a complex dynamic non-linear process that involves a set of non-linear markers and multi-variable interactions. Conventional statistical methods such as logistic regression (LR) [1] have been applied in cancer research, but did not always, if at all, result in reliable conclusions as far as an individual patient's prognosis of disease development is concerned [2]. Consequently, the development of intelligent methods that are able to yield reliable and accurate cancer prognosis has become important.

Artificial Neural Networks (ANNs), which are simplified mathematical models of the central nervous system, have been shown to be more capable of learning and modelling non-linear and complex systems than classical techniques [3]. Among various ANN architectures, multilayer backpropagation neural networks (MLBPNN) have been used widely for oncological prognosis and diagnosis [1].

Since Zadeh introduced fuzzy set theory in 1965 [4], it has been successfully applied in many areas including medicine. There is limited literature on fuzzy-based modelling in cancer, most of which having dealt with image analysis [5]. The fuzzy *K*-nearest neighbour algorithm (FK-NN) is one of the fuzzy-based classifiers that have been shown to be a powerful method for classification of patterns within data sets [6].

In recent years, several prognostic markers have been used as indicators of disease progression in breast cancer. Recently published research by Naguib et al. assessed four markers {DNA ploidy, S-phase fraction (SPF), G_0G_1/G_2M ratio, and nuclear pleomorphism index (NPI)} using artificial neural networks, and showed their influence on prediction of disease development [7]. They analysed them excluding the factors one by one from the data set, and made comments with respect to the predictive accuracy of the models. However, different combinations of the factors were not considered.

Image cytometry of breast cancer tissue usually involves the measurements of DNA ploidy, G_0G_1/G_2M ratio, SPF (%), and NPI. G_0G_1/G_2M ratio is the relative proportion of cells in the G_2M and G_0G_1 phases of the cell cycle. NPI is determined by two values: maximum (end) NPI (E-NPI) that shows normal nuclear shape and minimum (start) NPI (S-NPI) that indicates greater deviation from normality. In this study, those values of NPI have been considered as two separate markers rather than one marker as in [7]. Furthermore, in this study, two more factors {tumour histology and tumour grade} are included in the data set to assess their influence on prediction, as well as the relationship between combinations of the factors.

In previous studies conducted by our group, it was demonstrated that the fuzzy *K*-nearest neighbour (FK-NN) algorithm was capable of predicting prognosis in prostate and breast cancers, and a fuzzy measurement derived from the FK-NN was used as an indication of the importance of the prognostic factor subsets on prediction of the cancers [8 - 10].

In this study, the degree of importance of seven breast cancer prognostic factors will be analysed using a fuzzy measurement derived from the FK-NN and predictive accuracy of FK-NN, and compared with the results obtained by using the LR and MLBPNN. The paper is organised as follows: Section II deals with the structure of the breast

Report Documentation Page

Report Date	Report Type	Dates Covered (from... to)
25 Oct 2001	N/A	-

Title and Subtitle Prognostic Comparison of Statistical, Neural and Fuzzy Methods of Analysis of Breast Cancer Image Cytometric Data	**Contract Number**
	Grant Number
	Program Element Number
Author(s)	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) BIOCORE, School of Mathematical and Information Sciences, Coventry University Coventry, UK	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es) US Army Research, Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 09499-1500	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)

Distribution/Availability Statement
Approved for public release, distribution unlimited

Supplementary Notes
Papers from 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, October 25-26, 2001 held in Istanbul, Turkey. See also ADM001351 for entire conference on cd-rom.

Abstract

Subject Terms

Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UU

Number of Pages
4

cancer data set used in this study; Section III explains the methods; Section IV provides experimental results and Section V discusses the results. Conclusions are given in Section VI.

II. THE BREAST CANCER DATA STRUCTURE

The data set was collected from image cytometric data of fine-needle breast aspirates for 100 women who were clinically diagnosed with breast disease in the form of carcinoma or benign conditions. It consists of 7 prognostic markers measured for each patient. These are listed in Table I and used as input parameters to the different methods.

In this analysis, two corresponding outputs for histological assessment were predicted: alive or dead (within 5 years of diagnosis). In the assessment of the 100 patients, 86 patients had survived for 5 years with or without the disease, whereas 14 patients had died within 5 years of diagnosis.

III. METHODS

A. Logistic Regression

The goal of an analysis using the logistic regression (LR) method is the same as that of any model building technique used in statistics, that is to find the best fitting and most parsimonious model to describe the relationship between an outcome (dependent or response variable) and a set of independent (predictor or explanatory) variables. The outcome variable in logistic regression is binary or dichotomous, i.e. "yes" or "no", which is a feature that distinguishes the LR method from other statistical regression methods [11].

There have been several methods used for estimation and optimisation of the LR. The most commonly used method of estimation is maximum likelihood, where the Newton-Raphson algorithm is used to effectively estimate and optimise LR parameters [12].

In LR-based statistical analysis, univariate and multivariable factor analyses are used to assess the factors to determine their importance on predictive accuracy. Univariate analysis considers only one variable at a time, whereas multivariable analysis is applied to more than one variable or a whole variable set. A widely adapted approach to multivariable analysis for feature selection using LR is that of a stepwise method in which variables are selected either for inclusion or exclusion from the model in a sequential fashion based on statistical criteria. There are two main versions of the stepwise procedure: (a) forward selection with a test for backward elimination, and (b) backward elimination followed by a test for forward selection. In the stepwise method, errors are assumed to follow a binomial distribution, and significance is assessed via the likelihood ratio χ^2 test [11]. Thus, at any step in the procedure, the most important variable will be the one that produces the greatest change in the log-likelihood relative to a model not containing the variable.

TABLE I

Breast Cancer Prognostic Factors	Index
Tumour Histology Type	1
Tumour Grade	2
DNA Ploidy	3
SPF (%)	4
G_0G_1/G_2M Ratio	5
S-NPI	6
E-NPI	7

B. Multilayer Backpropagation Neural Networks

Multilayer feedforward neural networks (MLFFNN) have gained their popularity due to their capability of performing arbitrary mappings. Such mappings are possible if a sufficient number of hidden units are provided and if the network can be trained, that is if a set of weights that perform the desired mapping can be found [3].

Backpropagation (BP) is a general supervised method for iteratively calculating the MLFFNN's weights and biases. This type of MLFFNN is termed multilayer backpropagation neural networks (MLBPNN). It uses a steepest descent technique that is very stable when a small learning rate is selected, but has slow convergence properties [3]. Several methods for speeding up BP such as momentum and variable learning rates have been proposed [3]. However, the general method used for applications of MLBPNN is a steepest descent method with learning and momentum terms [1].

C. Fuzzy K-Nearest Neighbour Classifier

The fuzzy *K*-nearest neighbour algorithm (FK-NN), which is one of the widely used fuzzy based pattern classification methods, was proposed by Keller at al. in 1985 [6], and has been shown to be a powerful pattern classifier with various fields of application including medicine [6, 8, 9, 10]. This method is a function of class membership degrees and distances between a pattern to be classified and patterns of which class membership degrees are previously known. A class membership degree between 1 and 0 is computed using the first *K* (the number of neighbourhoods) minimum distances and the class membership degrees.

The FK-NN gives not only a class to which the pattern is assigned but also the class membership degree that provides information about the certainty of the classification decision. For example, if a pattern's membership degrees of classes 1 and 2 are 0.99 and 0.01, respectively, one can obviously be convinced that class 1 is the class to which the pattern belongs. However, if a pattern's membership degrees of classes 1 and 2 are 0.51 and 0.49, a high degree of confusion

arises as to which class it should be assigned. We refer readers to [6, 8, 9, 10] for details of the algorithms.

D. Fuzzy Measurement

The fuzzy measurement is a function of a class membership computed by means of the FK-NN algorithm. This measurement gives a degree of importance between 0 and 1 for each class (e.g. class-*c*) for subsets of the factors showing a degree of how representative the subsets can be of class-*c*. The highest value of the measurement is considered as a degree of importance of the subset. We refer readers to [8, 9, 10] for details of the algorithms.

E. Leave-One-Out Method

The "leave-one-out" method, regarded as a more reliable assessment of a classifier performance [11], will be used for LR, MLBPNN and FK-NN in order to test every pattern within the data sets. This method does not require dividing the data set into learning and testing data sets. The "leave-one-out" method leaves one pattern out of the learning data set and uses it as a test pattern every time. This procedure continues until each pattern is tested. This method avoids confusion over randomly selecting learning and test data sets. The method is also particularly suitable when a small amount of data is available, which is very common in oncological prognostic prediction applications. The results of testing each pattern by using the LR, MLBPNN and FK-NN are used to determine predictive accuracy.

IV. EXPERIMENTAL RESULTS

The following provides the experimental results obtained using LR, MLBPNN and FK-NN methods, respectively, for breast cancer prognosis of the image cytometric data.

Throughout these analyses, 99 combinations (subsets) of the prognostic factors from the 7-factor model to the 3-factor model were examined to determine the best subset(s) of the factors with the highest predictive accuracy and to interpret the interactions between the factors.

A. Logistic Regression (LR)

For the LR-based statistical analyses, the entire data set was first analysed by using SPSS [13]. In order to obtain an optimal model, the backward stepwise method was employed. A stratified model was {tumour histology, tumour grade, DNA ploidy, SPF (%), E-NPI}. This model was also used to obtain the leave-one-out predictive accuracy. The results are listed in Table II, where Class-1 and Class-0 refer to survival with or without the disease, or death within 5 years of diagnosis, respectively.

TABLE II
PREDICTIVE ACCURACY (%) OBTAINED BY USING LR

Methods	Class - 1	Class - 0	Total
Entire data set analyses by means of SPSS	97.90	14.30	86.00
Leave-one-out	95.35	00.00	82.00

B. Multilayer Backpropagation Neural Networks (MLBPNN)

For MLBPNN analysis, the neural network toolbox of MATLAB [14] was used. The network structure consisted of three-layers with ten neurons in the hidden layer. It was trained for 5000 training cycles. The highest predictive accuracy for the leave-one-out experiments was 87% obtained from two subsets: {tumour histology, DNA ploidy, SPF (%), G_0G_1/G_2M ratio, S-NPI, E-NPI} and {tumour histology, DNA ploidy, SPF (%), E-NPI}.

C. Fuzzy K-Nearest Neighbour Classifier (FK-NN)

The FK-NN analyses were carried out for up to eight neighbourhoods, i.e., K=1 to 8. The highest predictive accuracy obtained from K=6 and 7 was 88%. The eight different subsets yielded the same predictive accuracy of 88%. Among these subsets, four had the highest fuzzy measurement of 0.8183, and should be considered as the most important subsets of the factors since the fuzzy measurement denotes a degree of importance as mentioned in section III-D. They are 1: {tumour histology, grade, DNA ploidy, SPF (%), G_0G_1/G_2M ratio, S-NPI, E-NPI}, 2: {tumour histology, grade, DNA ploidy, SPF (%), G_0G_1/G_2M ratio, E-NPI}, 3: {tumour histology, grade, DNA ploidy, SPF (%), G_0G_1/G_2M ratio, S-NPI}, 4: {tumour histology, grade, DNA ploidy, SPF (%), G_0G_1/G_2M ratio}.

V. DISCUSSION

Prediction of breast cancer prognosis was carried out by means of three different methods. The predictive accuracy obtained from these methods and the corresponding stratified models are presented.

The LR statistical method of analysis identified the 5-factor model {tumour histology, grade, DNA ploidy, SPF (%), E-NPI}. However, this method highlights the difference in predictive accuracy between the leave-one-out technique, which refers to an individual patient's prognostic prediction, and the entire data set, these being 82% and 86%, respectively. It thus reveals that this method must be validated on data not used for its design, and that the result may not be reliable.

MLBPNN yielded a predictive accuracy of 87%, which is higher than that of LR. MLBPNN identified two models of the following 6 and 4 factor subsets: {tumour histology, DNA ploidy, SPF (%), G_0G_1/G_2M ratio, S-NPI, E-NPI} and {tumour histology, DNA ploidy, SPF (%), E-NPI}.

The FK-NN method identified the eight different subsets with the same highest predictive accuracy of 88%. In order to precisely identify the most significant subset, we selected the four subsets that yielded the highest value of the fuzzy measurement. A closer analysis of these subsets shows that the subset {tumour histology, DNA ploidy, SPF (%), G_0G_1/G_2M ratio} has the least number of prognostic factors, thus demonstrating that the excluded factors did not enhance the predictive accuracy and thus do not lend any significant contribution to the predictive accuracy of breast cancer prognosis of image cytometric data.

LR and MLBPNN results showed that maximum (end) NPI (E-NPI) is included in the model and is thus a more important factor than minimum (start) NPI (S-NPI). It should therefore be noted that E-NPI and S-NPI should be considered as separate factors, unlike the results reported in [7].

Furthermore, the three prognostic factors {tumour histology, DNA ploidy, and SPF} that are common in the models identified by all three methods may be the most significant factors for breast cancer prognostic prediction of image cytometric data. It is also shown that the additional factor {tumour histology} has been found to be significant in order to attain the highest predictive accuracy.

VI. CONCLUSIONS

The results for prognostic prediction of breast cancer patients by means of LR as a conventional statistical method, MLBPNN as a neural network method, FK-NN as a fuzzy logic method have been presented. The highest predictive accuracy is obtained from the FK-NN approach, while LR yielded the poorest predictive accuracy. We are of the opinion that the FK-NN and the fuzzy measurement are the best for achieving higher predictive accuracy and determining the most important prognostic factor(s) and subset(s) of these factors. In addition, the results suggest that the three prognostic factors {tumour histology, DNA ploidy and SPF} may be the most important factors for accurate and reliable breast cancer prognostic prediction.

REFERENCES

[1] E. Biganzoli, P. Boracchi, L. Mariani, and E. Marubini, "Feed forward neural networks for the analysis of censored survival data: A practical logistic regression approach", Statistics in Medicine, vol.17 (10), pp.1169-1186, 1998.

[2] R.N.G. Naguib, M.C. Robinson, I. Apakama, D.E. Neal, and F.C. Hamdy, "Neural Network Analysis of Prognostic Markers in Prostate Cancer", Br. J. Urol., vol.77 (1), p.50, 1996.

[3] S. Haykin, *Neural Networks*, 2nd ed., Prentice Hall, 1999.

[4] L.A. Zadeh, "Fuzzy Sets", Information and Control, vol.8, pp.338-353, 1965.

[5] T.C. Cahoon, M.A. Sutton, and J.C. Bezdek, "Breast cancer detection using image processing techniques", Proceedings of the 9th IEEE Conference on Fuzzy Systems, vol.2, pp. 973-976, 2000.

[6] J.M. Keller, M.R. Gray and J.A. Givens, "A Fuzzy *K*-Nearest Neighbour Algorithm", IEEE Trans. on Systems, Man and Cybernetics, vol.15(4), pp.580-585, 1985.

[7] R.N.G. Naguib, H.A. Mat-Sakim, M.S. Lakshmi, V. Wadehra, T.W.J. Lennard, J. Bhatavdekar and G.V. Sherbet, "DNA Ploidy and Cell Cycle Distribution of Breast Cancer Aspirate Cells Measured by Image Cytometry and Analysed by Artificial Neural Networks for their Prognostic Significance", IEEE Trans. on Information Technology in Biomedicine, vol.3 (1), pp.61-69, 1999.

[8] H. Seker, MO. Odetayo, D. Petrovic, R.N.G. Naguib, and F. Hamdy, "Ranking Prostate Cancer Prognostic Markers Using A Fuzzy *K*-Nearest Neighbour Algorithm", Proc. of the World Congress on Medical Physics and Biomedical Engineering, July 2000, Chicago, USA (Abstract no: TH-Aa325-03).

[9] H. Seker, MO. Odetayo, D. Petrovic, R.N.G. Naguib and F. Hamdy, "A Soft Measurement Technique for Searching Significant Subsets of Prostate Cancer Prognostic Markers", Proceedings of the European Symposium on Computational Intelligence, August 30-September 1, 2000, Kosice, Slovakia, pp: 325-328.

[10] H. Seker, M. Odetayo, D. Petrovic, R.N.G. Naguib, C. Bartoli, L. Alasio, MS. Lakshmi and GV. Sherbet, "A fuzzy measurement-based assessment of breast cancer prognostic markers", Proceedings of the IEEE EMBS International Conference on Information Technology Applications in Biomedicine, 9-10 November 2000, Washington, USA, pp: 174 –178.

[11] DW. Hoswer and S. Lemeshow, *Applied Logistic Regression*, John Wiley & Sons, 1989.

[12] B. Flury, *A First Course in Multivariate Statistics*, Springer, 1997.

[13] SPSS for Windows, Release 10.0.5, SPSS Inc., Nov. 1999.

[14] H. Demuth and M. Beale, *Neural Network Toolbox for use with MATLAB: User's Guide, version:3*, The Math Works Inc., 1998.