Statistical and Soft Feature Evaluation Indices for Prostate Cancer Prognostic Factor Assessment

Huseyin Seker[1], Michael O. Odetayo[1], Dobrila Petrovic[1], Raouf Naguib[1], and Freddie Hamdy[2]

[1] BIOCORE, School of Mathematical & Information Sciences, Coventry University, Coventry, UK

(h.seker, m.o.odetayo, d.petrovic, r.naguib@coventry.ac.uk)

[2] Royal Hallamshire Hospital and Sheffield University, UK

Abstract

In this paper, statistical, artificial neural networks and fuzzy based feature evaluation indices are analysed in order to determine the importance of prostate cancer prognostic markers. Seven prognostic markers are assessed in terms of 3 output classes using logistic regression as a statistical method, multilayer feedforward back propagation neural networks (MLFFBPNN) as a neural network tool, and fuzzy *k*-nearest neighbour algorithm (FK-NN) as a fuzzy method. The efficiency of the MLFFBPNN and FK-NN based indices is shown where the statistically based one fails to identify a clinically significant factor.

Keywords: Survival analysis, oncology, fuzzy *k*-nearest neighbour, multilayer feedforward backpropagation neural networks, logistic regression.

1 Introduction

Prognosis of prostate cancer is a complex dynamic non-linear process that involves a set of non-linear factors and multi-variable interactions. Conventional statistical methods such as logistic regression (LR) [1] have been applied in predicting cancer prognosis, but did not always, if at all, result in reliable conclusions as far as an individual patient's prognosis of disease development is concerned [6]. Such statistically based models employ the forward and backward "step-wise" methods in selecting important markers from a given set of data, and therefore can result in different models that may not be optimal due to the correlation and the interaction existing between the markers. In stepwise logistic regression, errors are assumed to follow a binomial distribution, and significance is assessed via the likelihood ratio χ^2 test. 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 [4].

Machine learning methods such as artificial neural networks, and in particular multilayer feedforward backpropagation neural networks (MLFFBPNN), have been applied to cancer prognosis and in determining the most/least important factors [1]. Recently, a fuzzy based method has been proposed for the same aim [7]. It has been shown that more than one subset of the prognostic factors can yield the same predictive accuracy. In such cases, it is difficult to make a decision on which subset could be the most/least important one. It has therefore become necessary to use another measurement in addition to the predictive accuracy for a factor, or subset of the factors, to determine their respective influence on prognosis more precisely.

In this paper, we present a fuzzy and MLFFBPNN based feature and subset evaluation indices for determining the most/least important factor(s) and subset(s) of the factors for prostate cancer prognosis. In addition, the results are compared with a traditional statistical method (LR).

2 MLFFBPNN Feature Evaluation Index

MLFFBPNN based feature evaluation index has been proposed by De et al. [2]. The index is calculated based on the concept that the effect of a missing feature on the output of a trained network will depend on the importance of that feature. In fact, the more significant a feature is, the more pronounced will be its impact on the output of the network, and hence, the higher the feature index value will be. De at al. used the entire data set to train the network, but did not select any test data set

to validate the network in order to test its generalisation ability for reliable feature selection.

3 A Fuzzy Feature Evaluation Index

A fuzzy based feature evaluation index has been proposed and used for prognostic factor assessments by Seker et al. [7]. This index is a function of class memberships (μ) of the patterns to be classified, computed by means of the fuzzy k-nearest neighbour algorithm [5], and the actual class memberships (u) of the patterns, which were determined a-priori. The index can be formulated as follows:

$$W^f = \frac{\sum_{i=1}^{N} u(i).\mu^f(i)}{\sum_{i=1}^{N} u(i)}$$

where N is the number of patterns within the data set, f indicates a subset of the factors, and W^f is the fuzzy feature evaluation index for subset f, which can also be regarded as "a weighted average class membership". Similarly to pattern class memberships, this measurement gives a degree of importance between 0 and 1 for the subsets of the factors, indicating how significant the subset is for prognosis. A subset that yields the highest value of the index is considered as the most important one.

The index can be used together with predictive accuracy, as a secondary measurement, to precisely identify the most/least significant factor(s) and/or subset(s) of the features.

4 Prostate Cancer Patients and Data Structure

Prognostic analysis is a function of a set of prognostic factors collected for each cancer patient. A reliable and accurate prediction depends inherently on these factors.

In recent years, several factors have been identified that have been regarded as indicators of prostate cancer progression. Tumour stage and volume, serum prognostic antigen levels, histopathological grading and DNA tumour ploidy status have been shown to correlate with prognosis and survival. In addition to clinical prognostic markers, several new factors are emerging which may have a varying degree of significance in predicting clinical outcome. Amongst these novel experimental factors are the genes regulating programmed cell death, otherwise known as apoptosis. These include the tumour suppressor gene p53 and the proto-oncogene bcl-2. It has been previously shown that the combination of bcl-2 overexpression and p53 nuclear accumulation by immunohistochemistry correlates strongly with hormone refractory prostate cancer [6].

For this study, four conventional and two experimental prostate cancer prognostic factors in addition to treatment information (listed in Table 1) were collected from 41 men with histologically proven prostate cancer, whose age ranged from 47 to 86 years (median 73 years). Twenty men (49%) had evidence of skeletal metastasis as demonstrated by technetium-99m isotope bone scanning, and received hormone manipulation. Eleven patients (27%) had clinically localised disease and received either "watchful waiting" or external beam irradiation. The remaining ten men (24%) had locally advanced cancers and received either radiotherapy or hormone manipulation. Follow-up ranged from 34 to 68 months (median 56 months). Of the patients, 5 had not responded to initial treatment, 20 developed resistant prostate cancer, and the remaining 16 patients were alive and well at the last follow-up period. Consequently, outcomes are categorised into 3 classes as listed in Table 2.

5 Results and Discussion

Feature evaluation indices and predictive accuracy are obtained using LR based backward stepwise method, MLFFBPNN and FK-NN. For MLFFBPNN and FK-NN based analyses, 99 combinations of the factors starting from a 7-marker model down to 3-marker model are analysed to determine the most/least important subsets of the factors.

SPSS [8] was used for the LR based analysis. The classes were analysed separately since the LR function allows the analysis of only one dependent variable at a time. Consequently, two classes at a time are aggregated in turn, e.g., classes-II and III are considered as one class during the analysis of class-I, and so on. For class-I, the identified model that yielded 87% predictive accuracy was {2, 5}. For class-II, the identified model that gave 80.5% predictive accuracy was {1, 2, 3, 7}. For class-III, the identified model that yielded 61% predictive accuracy was {4, 5}. The significance value of classes I and II was 0 while that of class-III was 0.032. The overall predictive accuracy of the models

obtained for all classes were 61%, 68.3%, and 65.9%, respectively.

Table 1: Prostate cancer prognostic factors

PROGNOSTIC FACTORS	INDEX
Conventional prognostic factors	
Tumour stage (T1-T4)	1
Skeletal metastasis (M0-M1)	2
Gleason score (2-10)	3
Serum PSA (1.2->2000)	4
Experimental prognostic factors	
p53 immunostaining (positive or negative)	5
bcl-2 immunostaining (positive or negative)	6
Treatment (hormonal, radical surgery or observation)	7

Table 2: Output classes for prostate cancer prognosis

CLASS	DESCRIPTION
I	No response to any type of treatment
II	Sustained complete response to treatment or no progression in treated patients (that is, the patient is alive and well at the last follow-up)
III	Relapse following initial successful treatment or disease progression in untreated patients.

For MLFFBPNN analysis, the neural network toolbox of MATLAB [3] was used. The network structure consisted of three-layers in which the hidden layer had ten neurons. The entire data set was trained for 5000 training cycles. Since MLFFBPNN sets its weights randomly, it was run 3 times and the average results are presented. The model that yielded the highest value for the MLFFBPNN based feature evaluation index of 39.6641 was {2,6,7} in which case the predictive accuracy was only 14.6%. This indicates that the factors {1,3,4,5} not included in this model are the most important factors. The model consisting of {1,2,3,4,5,6} gave the smallest value of the feature evaluation index (0.00001) in which case the predictive accuracy was 51.2%. It means that factor {7} which is not included in this model has no significant contribution to the predictive accuracy. It should also be noted that the index value of the models that exclude factor {4} ranges from 39.661 to 14.1682 while those of the models including factor {4} were not higher than 0.77. This MLFFBPNN based index suggests that factor {4} is the most significant prognostic factor. We reached the same conclusion when we analysed each of the classes individually. However, the LR based analyses identified {4} as one of the most important factors only for class-III while it is not for classes-I and II.

For the FK-NN based analysis, a MATLAB program was developed. The results were obtained for k=1,2, and 3 using the leave-one-out method. Among these results, k=2 yielded the highest predictive accuracy of 68.3%. For this case, the index value of 0.6202 was obtained from the model {2,3,4,7}. Similarly, for the results of the MLFFBPNN, when {4} was excluded from the model, the index value increased. For example, the index value for the models with {4} was not less than 0.5955, while those for the models without {4} were not higher than 0.4. The model {1,3,5,6} yielded the smallest index value of 0.20 and the poorest predictive accuracy of 36.6%. The FK-NN based index identified factor {7} as one of the important factors while the MLFFBPNN based index found that it did not enhance the predictive accuracy in the prostate cancer prognosis. The stratified subsets of the factors using all three methods studied are summarised in Table-3.

Factor {4: serum PSA} has been widely used for prostate cancer prognosis and regarded as a very reliable factor for monitoring the disease and the patient's response to therapy. PSA is also increasingly used for early detection and screening as widely discussed in the medical literature [9]. This marker was identified by the MLFFBPNN and FK-NN based indices, while the LR based analysis failed to identify it as a significant marker.

It should be noted that FK-NN based analysis was carried out using the leave-one-out method, whereas the entire data set was used for the LR and MLFFBPNN based analyses. A method's generalisation ability can be measured from the results obtained using a set of data not used in its design. The stratified marker subsets obtained using LR and MLFFBPNN were then subjected to the leave-one-out analysis. Predictive accuracy results for all methods are listed in Table-3. For LR, probability degrees of the patterns for each class were computed, and the patterns were assigned to a class that yields the highest degree of probability. For MLFFBPNN, the results were obtained from the

Table-3: Leave-one-out results for the most significant subsets identified using LR, MLFFBPNN, and FK-NN

Methods	The identified subset	Predictive accuracy (%)			
		Class-I	Class-II	Class-III	Total
LR	{2,5}	0	75	40	48.8
	{1,2,3,7}	0	68.8	60	56.1
	{4,7}	0	75	55	56.1
MLFFBPNN	{1,2,3,4,5,6}	0	6.3	100	51.2
FK-NN	{2,3,4,7}	40	75	70	68.3

most significant subset {1,2,3,4,5,6}, which was previously identified using its index. It can be seen from Table 3 that the highest predictive accuracy was obtained using FK-NN, which indicates that the generalisation ability of the identified model using FK-NN is better that LR and MLFFBPNN as far as an individual patient's prognosis is concerned.

6 Conclusions

We have presented the statistical, MLFFBPNN and FK-NN based soft feature evaluation indices for the assessment of prostate cancer prognostic factors. The efficiency of the MLFFBPNN and FK-NN based indices is shown where the statistical based index fails to identify a clinically significant factor. It should be noted that the results obtained show that different methods stratified different significant factors or subsets of factors, and therefore only one method's outcomes alone should not be relied upon for the prognostic purposes. In addition, the results suggest that the existing statistical methods may not be reliable as far as both identification of significant prognostic factors and prediction of disease development in the case of individual patients are concerned. It should also be pointed out that the FK-NN based index could be more reliable since it uses a set of data for test (validation) that is not used for its design, unlike in LR and MLFFBPNN analyses.

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.K. De, N.R. Pal, and S.K. Pal, "Feature analysis: neural network and fuzzy set theoretic approaches", Pattern Recognition, vol.30(10), pp:1579-1590, 1997.

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

[4] D.W. Hoswer and S. Lemeshow, *Applied Logistic Regression*, John Wiley & Sons, Inc., 1989.

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

[6] 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", British Journal of Urology, vol.77(1), pp:50, 1996.

[7] H. Seker, M.O. Odetayo, D. Petrovic, R.N.G. Naguib, C. Bartoli, L. Alasio, M.S. Lakshmi, and G.V. 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.

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

[9] U.H. Stenman, J. Leinonen, W.M Zhang, and P. Finne, "Prostate-specific antigen", Seminars in Cancer Biology, vol.9, pp:83-93, 1999.