AN INVESTIGATION OF THE EFFECT OF INPUT REPRESENTATION IN ANFIS MODELLING OF BREAST CANCER SURVIVAL

Hazlina Hamdan and Jonathan M. Garibaldi
Intelligent Modelling and Analysis (IMA) Research Group, School of Computer Science
The University of Nottingham, Jubilee Campus, Wollaton Road, Nottingham, NG8 1BB, U.K.

Keywords: Adaptive neuro-fuzzy inference system, Survival analysis, Breast cancer, Nottingham prognostic index.

Abstract: Fuzzy inference systems have been applied in recent years in various medical fields due to their ability to obtain good results featuring white-box models. Adaptive Neuro-Fuzzy Inference System (ANFIS), which combines adaptive neural network capabilities with the fuzzy logic qualitative approach, has been previously used in modelling survival of breast cancer patients based on patient groups derived from the Nottingham Prognostic Index (NPI), as discussed in our previous paper. In this paper, we extend our previous work to examine whether the ANFIS model can be trained to better match the data with the NPI variable represented as a real number, rather than a categorical group. Two input models have been developed and trained with different structures of ANFIS. The performance of these models, in the capability to predict the survival rate in survival of patients following operative surgery for breast cancer, is examined.

1 INTRODUCTION

The use of artificial intelligence (AI) techniques in the medical field in the early 1970s emerged to model expert behaviour by utilising the knowledge and representing it in symbolic form. Medical AI has since become very popular and has more recently been accepted by clinicians for its ability to produce high-quality results and demonstrate improvements upon previous techniques used (Joseph and David, 2006). In clinical situations such as diagnosis, treatment and prognosis in which there are complex interactions of clinical, biological and pathological variables, computerised analytical tools are needed to exploit the relationships between these variables. Soft-computing approaches including artificial neural networks and fuzzy inference systems (and many others) have been used to address this problem.

An artificial neural network (ANN) is an information processing system inspired by the structure of the human brain. ANNs have been the subject of great interest, following the discovery of the back-propagation algorithm, and recently have become very popular in the prediction of survival in medical contexts (Joseph and David, 2006; Lisboa, 2002; Burke et al., 1997). With their ability to learn through experience, neural networks work by detecting patterns in data, learning from the relationships and adapting to them. This knowledge is then used to predict the outcome for new combinations of data.

Fuzzy inference is based on the concepts of fuzzy set theory, fuzzy if-then rules, and fuzzy reasoning in which a mapping from a given input to an output is defined based on expert knowledge. The knowledge is encoded as a set of explicit linguistic rules, which can be easily understood by people without technical expertise. In medical fields, fuzzy inference have been used extensively for over a decade in data classification, decision analysis, diagnosis and prognosis (Yardimci, 2009). As fuzzy information deals with knowledge that is uncertain, ambiguous or imprecise, it is suitable to be used in the medical contexts to represent certain elements as members of sets with some degree of membership.

A hybrid methodology which combines the advantages of ANNs and fuzzy inference known as adaptive neuro-fuzzy inference system (ANFIS) technique was presented in modelling survival (Hamdan and Garibaldi, 2010). In this previous study, the Nottingham Prognostic Index (NPI) variable was represented as a categorical group. We now present a further study in which we extend our previous work by examining whether the ANFIS model can be trained to better match the data with the NPI variable repre-

Hamdan H. and M. Garibaldi J..
AN INVESTIGATION OF THE EFFECT OF INPUT REPRESENTATION IN ANFIS MODELLING OF BREAST CANCER SURVIVAL.
DOI: 10.5220/0003081100990104
In *Proceedings of the International Conference on Fuzzy Computation and 2nd International Conference on Neural Computation* (ICFC-2010), pages 99-104
ISBN: 978-989-8425-32-4

sented as a real number. In addition, two input models are presented to the ANFIS model and comparison of these two models is made as to the effect of membership function and generalization performance.

2 BACKGROUND

2.1 Breast Cancer

Cancer is a leading cause of death worldwide, as reported by the World Health Organization (WHO, 2010). Lung, stomach, liver, colon and breast cancer are all major contributors to the overall cancer mortality each year. Breast cancer is one of the most common cancers to afflict the female population. It is estimated that one in nine women in the UK will develop breast cancer at some point in their life (Cancer Research UK, 2010).

Breast cancer is a malignant tumour that develops from uncontrolled growth of cells in the breast. A malignant tumour is composed of cells that invade or spread to other parts of the body. The exact cause of the breast cancer is not really known, but is most likely to be a combination of genetic and environmental factors. However, in general, earlier diagnosis and treatment should increase the survival rates, as the disease is much easier to control if it has not spread to other parts of the body.

Breast cancer patients can be assigned into prognosis groups using a 'prognostic index'. The 'Nottingham Prognostic Index' (NPI) has been widely accepted in clinical practice to categorise patients into high (78%), intermediate (50%) or low (20%) risk groups. This index is based on pathological size, grade of tumor and the number of axillary nodes effected are identified significant in the prediction of survival (Galea et al., 1992). The NPI score can be calculated as:

NPI=0.2*pathological tumor size(cm)+lymph node stage+histological grade

Table 1 shows the accepted clinical cut-offs of the NPI score into categories patient into 'good', 'moderate' or 'poor'.

Table 1: Category of NPI score.

NPI score	**Category**
Less than 3.41	Good
Between 3.41 to 5.4	Moderate
Over 5.4	Poor

2.2 Survival Analysis

Survival analysis describes the analysis of data that corresponds to the time from when an individual enters a study until the occurrence of some particular event or end-point. In medical contexts, the event can be the response to a treatment, recurrence or disease-free survival, or death. An individuals with cancer cannot all be observed for the same length of time, because some individual are diagnosed at the beginning of the period under study, some near the end and others may be diagnosed at any time in the study.

Basically, survival data contains uncensored and censored observations. Uncensored observations involved patients who are observed until they reach the end of the study. Censored observations on the other hand, involve only patients who survive beyond the end or who are lost to follow-up at some point.

The survival function is defined as the probability that an individual survives longer than time t, where T denotes a positive random variable associated with the survival time, represented as (Biganzoli et al., 1998):

$$S(t) = P(T > t) \quad (1)$$

On the other hand, the hazard function, also known as conditional failure probability, is the probability an individual will die at a certain time t (conditioned on survival up to that time) and so denotes the instantaneous death rate. It can be shown in this form:

$$h_l = P(T \in A_l | T > t_{l-1}) = \left(\frac{S(t_{l-1}) - S(t_l)}{S(t_{l-1})} \right) \quad (2)$$

where the time interval $l = 1, 2, ..., L$ forms disjoint intervals $A_l = (t_{l-1}, t_l]$.

The survival and hazard functions are related to each other, in that the estimation of survival function can be written as:

$$S(t) = \prod_{l:t_l \leq t} (1 - h_l) \quad (3)$$

Statistical methods, such as the Kaplan-Meier estimate, are usually used to explain the data and to model the disease progression with the ability to handle censored data. A plot of the Kaplan-Meier is to represent the estimation of the survival function of some particular groups against time, can be view as a series of horizontal steps of declining magnitude.

2.3 ANFIS Architecture

The use of fuzzy logic in medical contexts may be said have been introduced by (Zadeh, 1969) in his paper entitled 'Biological application of the theory of

fuzzy sets and system' (Yardimci, 2009; **?**). Fuzzy logic is based on fuzzy sets that use linguistic variables with certain degree of membership and which can then be connected using IF-THEN rules to form a series of fuzzy rules. Fuzzy rules can have multiple antecedents connected with AND or OR operators, where all parts are calculated simultaneously and resolved into a single number. Consequents can also be comprised of multiple parts, which are then aggregated into a single output of a fuzzy set (Negnevitsky, 2005).

Fuzzy inference is a process of mapping from a given input to an output using the methods of fuzzy set manipulations. Two types of fuzzy inference most commonly used are the Mamdani method (Mamdani and Assilian, 1975) and the Sugeno method (Sugeno, 1985). The difference between these two fuzzy inferences methodologies is the specification of the consequent part. In the Mamdani method, consequents are fuzzy sets, and the final crisp output of Mamdani method is based on defuzzification of the overall fuzzy output using various types of defuzzification method. In contrast, in the Sugeno method, consequents are real numbers, which can be either linear or constant (zero-order Sugeno model). The final output (known as a singleton output membership function), is the weighted average of each rule's output.

Using an adaptation of the Sugeno fuzzy inference method, (Jang, 1993) proposed the adaptive neuro-fuzzy inference system (ANFIS) method that combined the neural network adaptive capabilities and the fuzzy logic qualitative approach. The ANFIS architecture contains a six-layer feed-forward neural network as shown in Figure 1 (Negnevitsky, 2005). Briefly, the functional of each layer are as given below:

Layer 1 is the input layer that passes external crisp signals to Layer 2.

Layer 2 known as the fuzzification layer, to determine the membership grades for each input implemented by the given fuzzy membership function.

Layer 3 is the rule layer, which calculates the firing strength of the rule as the product of the membership grades.

Layer 4 called the 'normalised firing strengths', in which each neuron in the layer receives inputs from all neurons in Layer 3, and calculates the ratio of the firing strength of a given rule to the sum of firing strengths of all rules.

Layer 5 is the defuzzification layer that yields the parameters of the consequent part of the rule.

Layer 6 is a single node that calculates the overall output as the summation of all incoming signals.

Full details of the ANFIS process can be found in (Jang, 1993) and (Negnevitsky, 2005).

ANFIS training can use alternative algorithms to reduce the error of the training. A hybrid approach, featuring a combination of the gradient descent algorithm and a least squares algorithm, is used for an effective search for the optimal parameters. The main benefit of such a hybrid approach is that it converges much faster, since it reduces the search space dimensions of the backpropagation method used in neural networks (Jang, 1993).

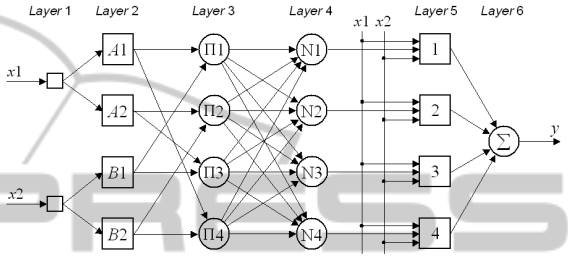

Figure 1: Adaptive Neuro-Fuzzy Inference System (ANFIS).

3 DATA AND METHODS

3.1 Data

A set of 958 breast cancer patients collected by the *Breast Cancer Pathology Research Group* in the University of Nottingham were used in a previous study to model the survival curve using the ANFIS model (Hamdan and Garibaldi, 2010). In the study, the patients are assigned into three groups of NPI whether good, moderate or poor (represented as 1, 2, and 3, respectively) based on the clinical cut-offs as shown in Table 1.

In this study, we used the same data set as a previous study with two variables as the input which are NPI values and survival time. However, the NPI variable in this study is presented as a real number (original values from the clinical). The ANFIS model was applied to the data set to examine whether the membership functions of a real-valued NPI can be trained to better match the data.

3.2 Methods

Data pre-processing was based on that of the non-linear method known as Partial Logistic Artificial Neural Network (PLANN) (Biganzoli et al., 1998) to produce smooth estimation of hazard rate. This method was created to allow the use of a standard

back-propagation ANN architecture to be used for modelling survival. A major process is to perform a specific form of data replication that was used in training phase of ANFIS method.

As stated in a previous study (Hamdan and Garibaldi, 2010), for training purposes, each patient is replicated for all the intervals in which the patient is observed, using the event indicator as the target. The input of the network (survival time and NPI values) is replicated into t times which is the maximum survival time of an individual patient. The event attribute as a target of the network is also replicated and assigned as zero until the last time value is reached, where the event is 1 for occurrence and zero for censored. An example of replication as shown in Table 2 which suitable to be train by the ANFIS.

While for the testing data to find the estimation of hazard rate for each interval time, each patient is replicated until the maximum time is observed or the full study time is reached. The hazard rate for each interval is the mean of hazard rate of all patient in that particular interval, and this depends on the cut-off of NPI to group the patients into good, moderate or poor. The estimation of survival function is determined using equation (3).

Initial parameters of the fuzzy inference system have to be established before the training process commences. Several ANFIS model were configured with different numbers of membership functions for the survival time, ranging from 3 to 7 and the number of membership functions for the NPI variable is based on the clinical groups (which is three). Gaussians were used for the membership functions and constants were used for the rule outputs (a zeroth-order Sugeno model). Hybrid learning, the combination of gradient descent and least squares algorithm, was selected as the learning algorithm.

4 EXPERIMENTAL RESULTS AND DISCUSSION

Data from 958 breast cancer patients were subjected to the pre-processing described above before being passed to the training process. This section presents the result of two models input into the ANFIS model: the final membership function generated, the learning rate and the conditional event probability will be discussed. Also, a comparison of survival rate of two input models is made according to the Kaplan-Meier method.

Two models of inputs are presented to the ANFIS model. In the first input model, the survival time is based in months with an observation time of 120 months while, in the second input model, the survival time was transformed into a yearly basis, corresponding to a ten year period of observation. Both models used real values of NPI.

Table 2: Replication for training.

	Time interval	NPI	Event
Patient 1	1	4.4	0
	2	4.4	0
	3	4.4	1
Patient 2	1	2.8	0
	2	2.8	0
	3	2.8	0
	4	2.8	0
	5	2.8	0
Patient 3	1	6.3	0
	2	6.3	1

Four membership functions of survival time were finally selected as it was observed that these provided a smooth conditional hazard function for the both input models. Figure 2 shows the initial membership functions for the first input model.

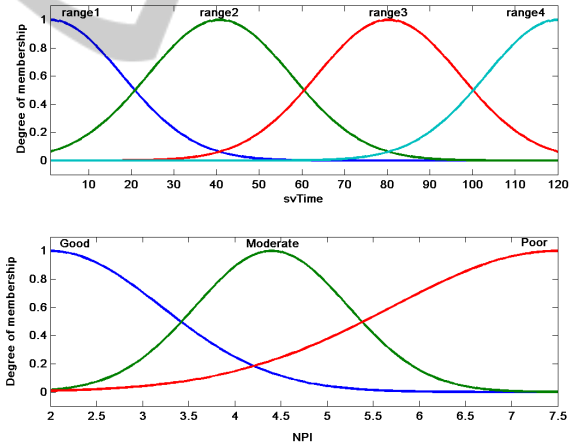

Figure 2: Initial membership functions of first input model.

In the training phase, the learning rate taken by the first input model is quite long, with 1100 epochs, rather than the second input model with only 100 epochs. In addition, the final membership functions of NPI generated by the second input model provide better interpretability than the first input model. Figure 3 and Figure 4 shows the final memberships of the first and second input models, respectively.

After both input models have been trained using the ANFIS methodology, we perform fuzzy inference calculations using the testing data as described in Sec-

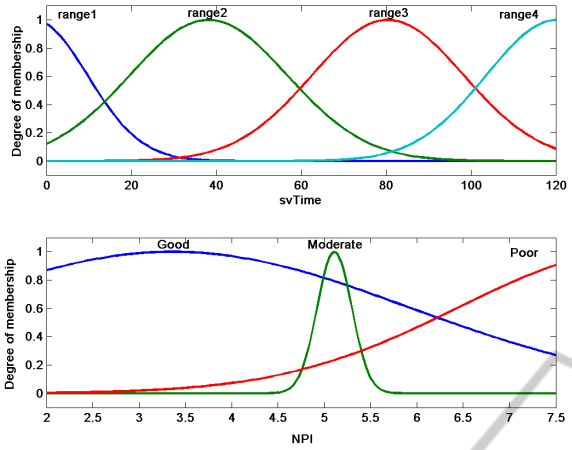

Figure 3: Final membership functions of first input model.

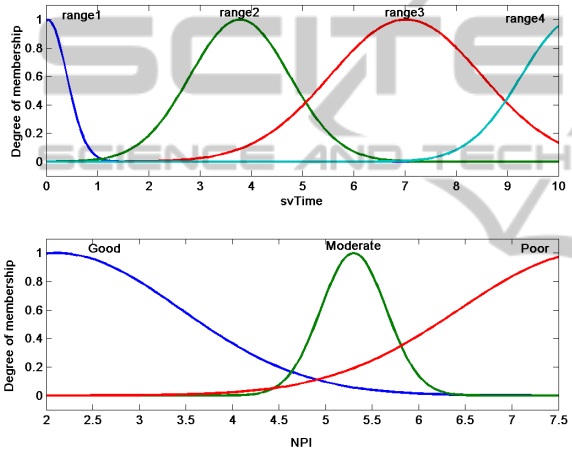

Figure 4: Final membership functions of second input model.

tion 3.2. The output of the testing is the estimation of conditional failure probability for each time interval (i.e. the hazard function). From this, the estimation of survival function using Equation (3) can be plotted. Figure 5 and Figure 6 shows the estimates of survival function for the first input model and second input model, respectively, against the Kaplan-Meier plot for the original (observed) data.

It can be seen that, while the fitted-curve obtained from the second input model are close to Kaplan-Meier plot for the 'poor' category (red line), the first input model produces better a fitted-curve for the 'moderate' category (green line). However, both input models gave approximately the same fitted-curve for the 'good' category. It can be seen that, the curves obtained from the ANFIS model are close to those of the Kaplan-Meier plot when the NPI value is presented as a real number.

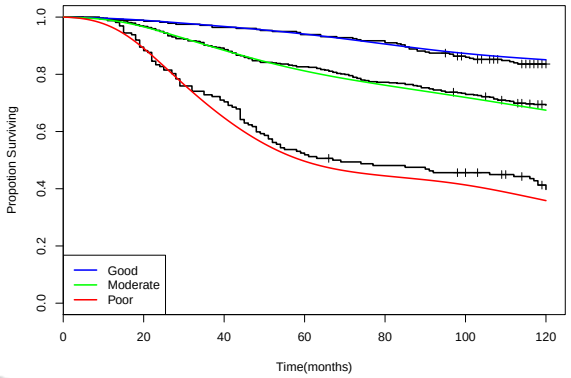

Figure 5: Survival curve of actual Kaplan-Meier (balck solid lines) estimated against the first input model (color lines).

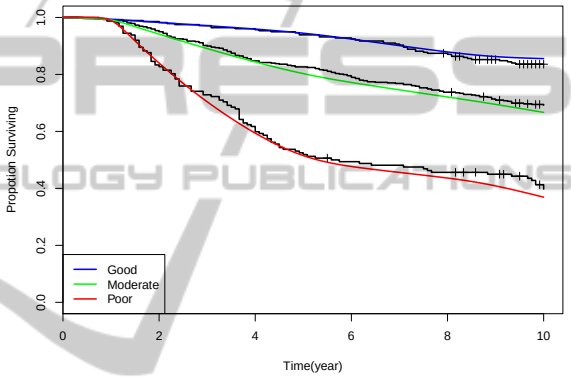

Figure 6: Survival curve of actual Kaplan-Meier (balck solid lines) estimated against the second input model (color lines).

5 CONCLUSIONS

The ANFIS models have been applied to the Nottingham Breast Cancer data set with the NPI variable represented as a real number to estimate the conditional failure probability and the survival curve. This compares to our previous work, in which the NPI was presented as a categorical input. Two input models have been developed and data replication performed in order for the data to used to train an ANFIS model. With the NPI represented in real values, the ANFIS model can estimate the proportional hazard rates and, furthermore, the survival function can be plotted.

In general, as ANFIS adapts the neural network learning, normally before training neural network, it is necessary to transform the data to new representation to reduce the dimensionality of input data and to optimise the generalization performance (Bishop, 2007). In our findings, when the input variables span similar ranges or scale, it produces a better interpretability on the final membership function with

short learning rate. That is, the results obtained when representing the time in months differ from those in which the time is represented in years, despite the fact this is just a simple scaling.

6 FUTURE WORK

In the future, we aim to investigate how to restrict the constant value of the singleton output of the rules producing by the ANFIS to be all positive, so that we can obtain a smooth curve of conditional probability with non-negative values in any of the time intervals.

Further investigations into the effects of scaling the inputs to the ANFIS model will also be undertaken, to see whether there are any significant effects on learning rate and/or final membership functions.

We also aim to create ANFIS models for other clinical data sets — we have recently obtained data for a cohort of over 400 colorectal cancer patients with ten year follow-up survival data.

ACKNOWLEDGEMENTS

The authors thank all members of the *Nottingham Breast Cancer Pathology Research Group*, and particularly Prof. Ian Ellis, Dr Andy Green and Dr Des Powe, for their help in preparing and providing the data set used in this study.

This study was supported by the Ministry of Higher Learning, Malaysia and Universiti Putra Malaysia (UPM).

REFERENCES

Biganzoli, E., Boracchi, P., Mariani, L., and Marubini, E. (1998). Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in Medicine*, 17(10):1169–1186.

Bishop, C. M. (2007). *Neural Networks for Pattern Recognition*. Oxford University Press, UK.

Burke, H., Goodman, P., Rosen, D., Henson, D., Weinstein, J., Harrell, F., Marks, J., Winchester, D., and Bostwick, D. (1997). Artificial neural network improve the accuracy of cancer survival prediction. *Cancer*, 79(4):857–862.

Cancer Research UK (2010). Uk breast cancer incidence statistics. Date last accessed: 18/05/2010.

Galea, M., Blamey, R., Elston, C., and Ellis, I. (1992). The nottingham prognostic index in primary breast cancer. *Breast Cancer Research and Treatment*, 22(3):207–219.

Hamdan, H. and Garibaldi, J. M. (2010). Adaptive neuro-fuzzy inference system (ANFIS) in modelling breast cancer survival. In *2010 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 573–580.

Jang, J.-S. (1993). Anfis adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man and Cybernetics*, 23(3):665–685.

Joseph, A. C. and David, S. W. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics*, 2:59–78.

Lisboa, P. J. G. (2002). A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural Networks*, 15(1):11–39.

Mamdani, E. H. and Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies,*, 7:1–13.

Negnevitsky, M. (2005). *Artificial Intelligence: a guide to intelligent systems*. Pearson Education Limited, Essex, England.

Ramesh, A. N., Kambhampati, C., Monson, J. R. T., and Drew, P. J. (2004). Artificial intelligence in medicine. *Annals of The Royal College of Surgeons of England*, 86:334–338.

Sugeno, M. (1985). *Industrial applications of fuzzy control*. Elsevier Science Pub. Co.

WHO (2010). Cancer. Date last accessed: 18/05/2010.

Yardimci, A. (2009). Soft computing in medicine. *Applied Soft Computing*, 9(3):1029 – 1043.

Zadeh, L. A. (1969). Biological application of the theory of fuzzy sets and systems. In *Proceeding of the International Symposium of Biocybernetics of the Central Nervous System*, pages 199–212.