

Application of Artificial Neural Network-Based Survival Analysis on Two Breast Cancer Datasets

Chih-Lin Chi^a, W. Nick Street^b, William H. Wolberg^c

^a Health Informatics Program, University of Iowa

^b Management Sciences Department, University of Iowa

^c Department of Surgery, University of Wisconsin

Abstract

This paper applies artificial neural networks (ANNs) to the survival analysis problem. Because ANNs can easily consider variable interactions and create a non-linear prediction model, they offer more flexible prediction of survival time than traditional methods. This study compares ANN results on two different breast cancer datasets, both of which use nuclear morphometric features. The results show that ANNs can successfully predict recurrence probability and separate patients with good (more than five years) and bad (less than five years) prognoses. Results are not as clear when the separation is done within subgroups such as lymph node positive or negative.

Keywords

Breast cancer prognosis, survival analysis, artificial neural networks, machine learning

Introduction

Survival analysis is the study of time to an event of interest, such as disease occurrence or death. The time to the event is called survival time, such as disease-free time. Prognosis is the prediction of the time to some future event, such as cancer recurrence. Therefore, we can see prognosis as a survival analysis problem [8]. This study uses an Artificial Neural Network (ANN) model for breast cancer prognosis, predicting how long after surgery we can expect the disease to recur. This decision support tool for prognosis can help to aid the doctor and patient in determining a course of post-operative treatment. To assist prognosis, an ANN model learns to predict the time to recur (recurrence time) from previous patients and then predicts outcome for the new patient. Some patients' time to recur are directly observed in the follow-up study. For others, we only know the time of their last check-up or disease-free survival time (DFS)

because these patients may change doctors, move away, or leave the study for other reasons. These patients are called *censored* cases.

Traditionally, analysis of censored data is performed using Cox proportional hazards regression [2]. In recent years, machine learning methods and particularly ANNs, have been widely used in prediction using censored data. This is because, as pointed out by Biganzoli et al. [1], feed-forward ANNs provide flexible non-linear modeling of censored survival data through the generalization of both discrete and continuous time models. This paper compares prediction results in two breast cancer datasets using a modification of Street's ANN model [12]. This model predicts probabilities of recurrence at different time intervals for each patient. By using a probability threshold, this model can differentiate patients with "good" or "bad" prognosis. We also show that the choice of training subsets can affect prediction results.

Background and Related Work

In survival analysis, Cox's proportional Hazards models [2] have been traditionally used to discover attributes that are relevant to survival, and predict outcomes. Smith et al. [11] transformed the output from Cox regression into survival estimation. However, the proportional hazards model is subject to a linear baseline. Cox regression makes two important assumptions about the hazard function: (1) Covariates affecting the hazard rate are independent, and (2) the ratio of risk in dying of two individuals is the same regardless of the time they have survived. De Laurentiis & Ravdin [3] suggested three situations in which artificial neural networks are better than Cox's regression model:

1. The proportionality of hazards assumption can not be applied to the data.
2. The relationship of variables to the outcome is complex and unknown.
3. There are interactions among variables.

These problems can be solved by non-linear models such as artificial neural networks. There are several approaches to the use of ANNs for survival analysis. For example, De Laurentiis & Ravdin [3], added a time input to the prognostic variables to predict the probability of recurrence. The original vector is transformed into a set of data vectors, one for each possible follow-up time. Before the recurrence time, the target value is set to 0, and to 1 at the time of event occurrence and all subsequent intervals. For censored cases, they used Kaplan Meier [6] analysis to modify the number of data points of non-survivors in each time interval. Biganzoli et al. [1] also treated the time interval as an input variable in a feed-forward network with logistic activation and entropy error function to predict the conditional probabilities of failure. Another form of artificial neural networks that have been applied to survival analysis is called “single time point models” [4]. In this model, a single time point t is fixed, and the network is trained to predict the t -year survival probability. This model can repeatedly predict several time points. For example, Kappen & Neijt [7] trained 6 artificial neural networks to predict survival of patients with ovarian cancer after 1, 2, ..., 6 years. Closely related to this approach are the so-called “multiple time point models”. In this kind of model, one neural model has K output units for K time points. The multiple time point model of Street [12] uses a single network to predict survival probability at each time point. He changed all the time points into survival probabilities as the target outputs, again using Kaplan Meier probabilities to set target values for censored points.

Data

The two breast cancer datasets used in this study were generated using the Xcyt image analysis program ([10] [14]). First, a sample of fluid is taken from the patient’s breast by a fine needle aspirate. An image of the fluid is transferred to a workstation by a video camera mounted on a microscope. Xcyt uses a curve-fitting program to determine the boundary of cell nuclei. Ten features are computed for each nucleus: radius, texture (variance of grey levels inside the boundary), perimeter, area, smoothness (local variation of radial segments), compactness, fractal dimension (of the boundary), symmetry, concavity, and the number of concave points. The mean value, extreme value, and standard error of these features for all cells are also

computed, resulting in a total of 30 morphometric features for each patient.

The first dataset is called the Wisconsin Prognostic Breast Cancer (WPBC) data and contains 151 censored cases and 47 recurrent cases. The second dataset is called the Love data [9] with 309 censored cases and 153 recurrent cases. There is a treatment status variable in the Love data and a lymph node status variable in the WPBC data. The treatment status variable indicates if a patient had hormone therapy or not. Lymph nodes status is the number of positive axillary lymph nodes, and is commonly used in cancer staging for prognostic purposes. Patients with one or more positive lymph nodes (designated WPBC+) have a generally poorer prognosis than those with no positive nodes (WPBC-). The date sets contain other prognostic factors such as tumor size, estrogen receptor status and progesterone receptor status which have been shown to be related with disease recurrence ([5] [13]). However, in order to compare two datasets, we used only nuclear morphometric features in this study.

Method

We used three-layer, feedforward ANNs with sigmoid activations in this work. The networks were trained using the standard backpropagation algorithm which contains three layers, an input layer, a hidden layer, and an output layer. The ANNs will learn the relationship between independent variables (each input node represents a variable in the input layer) and dependent variables (each output node represents a variable in the output layer). There are 30 nuclear features in the input layer. The number of hidden nodes is 20, and adaptive learning with 1000 epochs was used for training. In the output layer, each node represents a time interval, from six months to ten years, in intervals of six months. The training signal for each node represents the probability of disease-free survival for the example at that time point. For the recurrent cases, the network was trained with +1 for all output nodes up to the recurrence time, and 0 thereafter. For example, a patient with recurrence time between 3.5 and 4 years would have the target vector {1, 1, 1, 1, 1, 1, 1, 0}. For censored cases, we only know the disease-free survival time (DFS). The KM survival function is used to estimate the probability of recurrence for such cases at time points after the observed DFS time. The output nodes can be expressed as:

$$S_t = \begin{cases} 1, & 0 \leq t \leq DFS_i \\ S_{t-1}(1 - risk_t), & t > DFS_i \end{cases} \quad (1)$$

The *risk* of recurrence at time $t > 0$ is the conditional probability that a patient will recur at time t , given that they have not recurred up to time $t-1$. For example, consider an experiment containing a total of 20 patients. If two patients recurred in the first time interval, we have $risk(1) = 0.1$. Furthermore, two censored cases were observed in the first time interval, and two more recurrences were in the second interval. We have $risk(2) = 0.125$. A censored case with an observed DFS of 2.5 years may have an output vector of $\{1, 1, 1, 1, 1, 0.97, 0.94, 0.92, 0.91, 0.89, 0.89, 0.89, 0.79, 0.79, 0.79, 0.79, 0.79, 0.79, 0.79, 0.79\}$. The first five units are known disease-free survival probabilities, and the following time units are estimated from the KM survival function. We can consider this network to have been trained with survival probabilities, and the predicted outputs are survival probabilities for each time unit.

For the predicted output, we defined the first predicted output unit with an activation less than 0.5 as the predicted time to recur. For example, a predicted output of $[1, 0.95, 0.9, 0.85, 0.8, 0.75, 0.7, 0.65, 0.63, 0.6, \mathbf{0.48}, 0.43, 0.4, 0.37, 0.35, 0.3, 0.28, 0.2, 0.15, 0.13]$ corresponds to a predicted disease-free survival time of 5.5 years. A predicted disease-free time greater than five years is defined as a good prognosis, and less than five years is a bad prognosis. In our experiments, we perform ten-fold cross-validation, accumulate all the test predictions, and divide them into good and bad groups. The true disease-free survival of the resulting groups is plotted for visual inspection using Kaplan-Meier curves, and compared for significant differences using a Wilcoxon test. To be a useful prognostic tool, the actual outcome in the good prognostic group should be significantly better than those in bad group.

Results

Figure 1 compares the true Kaplan-Meier estimate of disease-free survival for the entire dataset with the predicted DFS rates accumulated from the test folds. The error bars are 95% confidence intervals from the predicted survival function. The error bars in the Love data is shorter than WPBC data because the sample size in Love data is twice that of WPBC data. Actual and predicted disease-free survival curves are

very close, and the shapes are very similar in both datasets, indicating that the overall estimated survival characteristics are accurate.

Fig. 2 shows the separation of good and bad outcomes is significant for both data (Wilcoxon test: $P = 0.0005$ in Love; $P = 0.0105$ in WPBC). This sort of stratification can help a clinician to determine a patient's probability of surviving free of recurrence for more than 5 years. Then the clinician can decide an appropriate treatment plan accordingly.

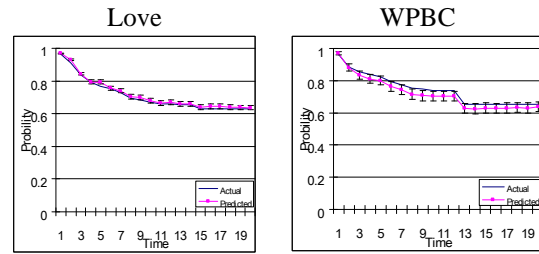


Figure 1: Actual survival compared to predicted survival curve.

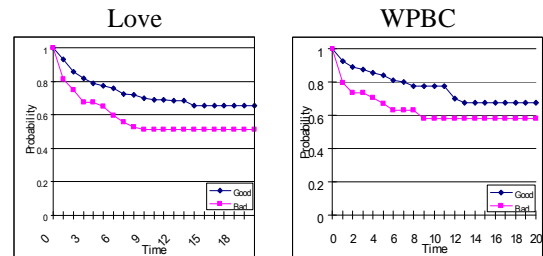


Figure 2: Disease-free survival curve of cases predicted to recur within five years (Bad) compared to those predicted to recur at some time greater than five years (Good).

In order to compare the separation result to a single time point model, we separated the training sets into good and bad outcomes using 5 years as the cut point and removed the censored cases less than 5 years in the training folds while performing the ten-fold cross-validation. Survival curves were then obtained using the true outcomes, separating the cases by the predicted class. We applied Naive Bayes (NB) and ANN to the single time point model with the two datasets. Interestingly, the single time point model separated the WPBC data well ($P=0.0021$ for NB; $P=0.0104$ for ANN using the Wilcoxon test on the true outcomes). However, the separation was inferior to our model for the Love data ($P=0.6159$ for NB; $P=0.1832$ for ANN).

If we divide a dataset into more homogeneous subgroups, such as Love treatment vs. Love control or WPBC- vs. WPBC+, our multi point predictive model is less effective. For example, in Figure 3, only the separation in the Love treatment group is significant ($P=0.0001$). To understand the reason, we investigate the separation crossing different subgroups. In the Love data, they are treatment good vs. treatment bad, treatment good vs. control bad, control good vs. treatment bad, and control good vs. control bad. In the WPBC data, they are WPBC- good vs. WPBC- bad, WPBC- good vs. WPBC+ bad, WPBC+ good vs. WPBC- bad, and WPBC+ good vs. WPBC+ bad. Table 1 and Table 2 show P values of the above comparison pairs in both datasets. These two tables show that the separation of the whole dataset may result from the separation crossing a different subgroup. For example, the significant separation in the whole Love dataset is not only from the separation between treatment good and treatment bad, but also from treatment good and control bad.

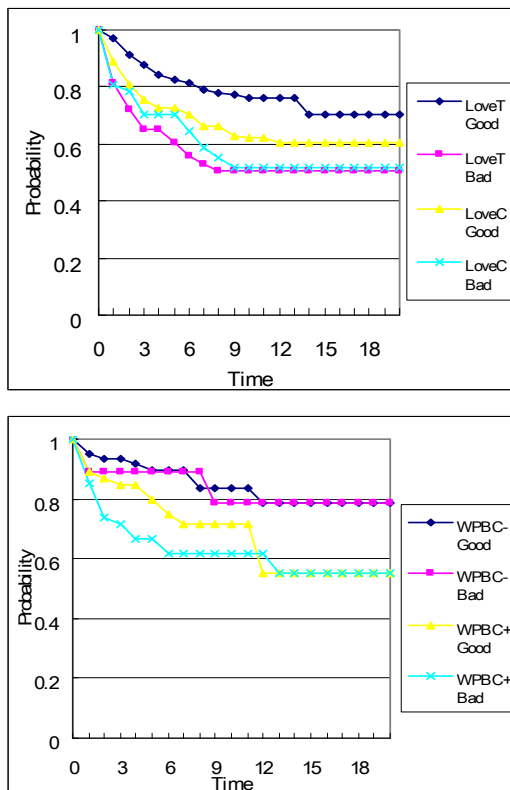


Figure 3: Both datasets were divided into two subgroups, and then separated by predicted outcome within each subgroup.

This result indicates that the prediction is good

enough to help discriminate good and bad prognostic groups from the patient population, but is not good at the same decision within a more homogeneous subgroup. There may be several reasons for the not-good-enough prediction. The first reason is the independent variables (only morphometric features) are not strong enough. Other strong predictors, such as tumor size, estrogen receptor status and progesterone receptor status, are not used in this experiment in order to compare two datasets. The second reason is the data size. The total data sizes of the WPBC and Love datasets are 198 and 462. For pattern recognition, the sizes of both datasets are small. Splitting each small dataset further into two subgroups will increase the difficulty of separation in each subgroup.

Table 1 P-values (Wilcoxon test) of separation in Love treatment and control

	Treat Good	Ctrl Good
Treat Bad	0.0001*	0.0794
Ctrl Bad	0.0015*	0.7523

Table 2 P-values (Wilcoxon test) of separation in WPBC- and WPBC+

	WPBC- Good	WPBC+ Good
WPBC- Bad	0.6349	0.4792
WPBC+ Bad	0.0025*	0.2590

Discussion

In this study, we used an ANN model to build a prognosis decision support tool and compare breast cancer prognosis results in two datasets. Our results show that this model can accurately predict the survival probability of each time period after a patient has a surgery.

In a clinical setting, prognosis is a very important indicator to determine a course of treatment, such as chemotherapy, after surgery. Often, patients are divided into high-risk and low-risk groups that follow different rules for therapy. Thus, a good classification is very important. In this research, we used recurrence at five years as a cut point to define the level of risk. The ANN model can answer the following question effectively: “After surgery, what is the level of recurrence risk for this breast cancer patient compared to all other patients?” Our results show that this ANN model can classify outcomes well in the whole dataset. This result is consistent with Street’s [12] original

study. However, if the prognosis question is, "What is the level of risk for a lymph -negative compared to all other lymph-negative patients?" this ANN model can not answer the question effectively. Using better predictive methods and larger data sets may improve the prediction.

We note that, for effective prognostic prediction, each model needs to be constructed on a set of comparable cases. For a number of reasons (including sample preparation and base population differences), the WPBC and Love data sets are not directly comparable. One final caveat is that the results presented here represent a cross-validated retrospective study. We have every reason to believe that results on new cases (drawn from the appropriate population) would be similar; however, no prospective study using this technique has been performed.

In future work, we can incorporate more prognostic factors such as tumor size, estrogen receptors status, progesterone receptors status, and HER2/neu status to improve the prediction.

References

- [1] 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, 1169–1186.
- [2] Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, 34: 187-220.
- [3] De Laurentiis, M. and Ravdin, P. M. (1994). A technique for using neural network analysis to perform survival analysis of censored data. *Cancer Letters*, 77:127-138.
- [4] De Laurentiis, M. and Ravdin, P. M. (1994). Survival analysis of censored data: neural network analysis detection of complex interactions between variables. *Breast Cancer Research Treatment*, 32:113-118
- [5] Gruvberger, S., Ringner, M., Chen, Y., Panavally, S., Saal, L. H., Borg, A., Ferno, M., Peterson, C., Meltzer, P. S. (2001). Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res* 2001, 61:5979-5984.
- [6] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53: 457-481.
- [7] Kappen, H. J. and Neijt, J. P. (1993). Neural network analysis to predict treatment outcome. *Annals of Oncology*, 4, Supplement: S31-34.
- [8] Lee, E. T. (1992) *Statistical Methods for Survival Data Analysis*. John Wiley and Sons, New York.
- [9] Love, R. R., Duc, N. B., Baumann, L. C., Anh P. T. H., To, T. V., Qian, Z., and Havighurst, T. C. (2004). Duration of signs and survival in premenopausal women with breast cancer. *Breast Cancer Research and treatment* 86: 117-124.
- [10] Mangasarian, O. L., Street, W. N., and Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43(4):570-577.
- [11] Smith, A. E. and Anand, S. S. (2000). Patient survival estimation with multiple variables: Adaptation of Cox's regression to give an individual's point prediction. In: *Proceedings of the IDAMAP*. Berlin, p. 47-55
- [12] Street, W. N. (1998). A neural network model for prognostic prediction. *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 540-546, San Francisco, 1998. Morgan Kaufmann.
- [13] West, M., Blanchette, C., Dressman, H., et al (2001). Predicting clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci USA* 98: 11462-11467
- [14] Wolberg, W. H., Street, W. N., and Mangasarian, O. L. (1994). Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. *Cancer Letters*, 77:163-171.

Address for correspondence

Chih-Lin Chi
Health Informatics Program
S283 John Pappajohn Business Building
The University of Iowa
Iowa City, IA 52242