Artificial Neural Networks in Identifying Areas with Homogeneous Survival Time

Małgorzata Krętowska[1] and Leon Bobrowski[1,2]

[1] Faculty of Computer Science, Białystok Technical University, Wiejska 45a, 15-351 Białystok, Poland
[2] Institute of Biocybernetics and Biomedical Engineering PAS, Ks. Trojdena 4, Warsaw, Poland
e-mail: mmac@ii.pb.bialystok.pl; leon.bobrowski@ibib.waw.pl

Abstract. In the paper an artificial neural network designed for prediction of survival time is presented. The method aims at identifying areas in the feature space homogeneous from the point of view of survival experience. The proposed method is based on minimization of a piecewise linear function. Appropriately designed dipolar criterion function is able to cope with censored data. Additional pruning phase prevents the network from over-fitting.

1 Introduction

Survival analysis is concerned mainly with prediction of failure occurrence. In medicine, the failure usually represents death or disease relapse. Failure time prediction is done on the base of datasets, which are gathered within the space of several years. The data, describing the analyzed phenomenon, contains patients characteristics (e.g. sex, age, outcomes of laboratory tests) and the failure time t. One of the most important and distinguishing features of survival data is censoring. For *uncensored* patients the failure occurred exactly at the time t (failure indicator $\sigma = 1$). *Censored* cases ($\sigma = 0$) contain incomplete information of the failure occurrence. The only information is that the failure did not occur before t.

Statistical methods used in analysis of survival data require some additional assumptions about the analyzed phenomenon. If the requirements are difficult to obey some other techniques are adopted. Among them artificial neural networks are considered as one of the most promising tools. Application of ANN in survival data analysis may be assess from the point of view of several criteria. One of the most common approach is based on the prediction of failure occurrence before a given time e.g. 5 years [6, 9]. Other authors divide survival time into several disjoint intervals and the neural network task is to choose the interval in which the failure is most likely to occur. The problem that appears in all proposed techniques is how to treat censored cases. The most popular idea is simply to omit them, another one is to impute the failure time using such methods as Cox regression models [7], decision trees [12] or Kaplan-Meier method [11]. The

neural network models that are able to cope with censored cases directly are a small part of all the proposed techniques. They are developed as a generalization of regression models and were used mainly to estimate the hazard function [2, 10].

In this paper we propose a neural network model aims at prediction of survival time. This is equivalent to identifying subgroups of patients with homogeneous response for a given treatment. The technique is able to deal with censored cases and do not require artificial division of survival time. The additional optimizing phase allows receiving the network with better generalization ability.

2 Learning phase

A neural network model, considered in the paper, consists of two layers: input and output layer. The output layer is built from neurons with binary activation function. From geometrical point of view a neuron divides feature space into two subspaces by using a hyper-plane $H(\mathbf{w}, \theta) = \{\mathbf{x} :< \mathbf{x}, \mathbf{w} >= \theta\}$, where $\mathbf{x}$ is a feature vector, $\mathbf{w}$ - a weight vector and θ is a threshold. The neuron is activated (the output $z = 1$) if the input vector $\mathbf{x}$ is situated on the positive site of the hyper-plane H. The layer of formal neurons divides N-dimensional feature space into disjoint areas. Each area is described by single output vector $\mathbf{z} = [z_1, z_2, \ldots, z_L]^T$, where $z_i \in \{0, 1\}$, L is the number of neurons in the layer. The network works correctly if all the areas are homogeneous from the point of view of analyzed problem. In case of survival data each area should include patients with similar survival times.

The proposed learning procedure is based on the concept of dipoles [3]. The dipole is a pair of different covariates vectors $(\mathbf{x}_i, \mathbf{x}_j)$ from the learning set. We distinguish mixed and pure dipoles. Mixed dipoles are formed between objects, which should be separated while pure ones between objects which are similar from the point of view of an analyzed criterion. In our approach pure dipoles are created between pairs of feature vectors, for which the difference of failure times is small, mixed dipoles - between pairs with distant failure times. Taking into account censored cases the following rules of dipole construction can be formulated:

1. a pair of feature vectors $(\mathbf{x}_i, \mathbf{x}_j)$ forms the pure dipole, if
   - $\sigma_i = \sigma_j = 1$ and $|t_i - t_j| < \eta$
2. a pair of feature vectors $(\mathbf{x}_i, \mathbf{x}_j)$ forms the mixed dipole, if
   - $\sigma_i = \sigma_j = 1$ and $|t_i - t_j| > \zeta$
   - $(\sigma_i = 0, \sigma_j = 1$ and $t_i - t_j > \zeta)$ or $(\sigma_i = 1, \sigma_j = 0$ and $t_j - t_i > \zeta)$

Parameters η and ζ are equal to quartiles of absolute values of differences between uncensored survival times. The parameter η is fixed as 0.2 quartile and ζ - 0.6.

We introduce two types of piece-wise linear and convex (CPL) penalty functions $\varphi_j^+(\mathbf{v})$ and $\varphi_j^-(\mathbf{v})$:

$$\varphi_j^+(\mathbf{v}) = \begin{cases} \delta_j - < \mathbf{v}, \mathbf{y}_j > & \text{if} < \mathbf{v}, \mathbf{y}_j > \leq \delta_j \\ 0 & \text{if} < \mathbf{v}, \mathbf{y}_j >> \delta_j \end{cases} \quad (1)$$

$$\varphi_j^-(\mathbf{v}) = \begin{cases} \delta_j + < \mathbf{v}, \mathbf{y}_j > & \text{if} < \mathbf{v}, \mathbf{y}_j > \geq -\delta_j \\ 0 & \text{if} < \mathbf{v}, \mathbf{y}_j > < -\delta_j \end{cases} \tag{2}$$

where δ_j is a margin ($\delta_j = 1$), $\mathbf{y}_j = [1, x_1, \ldots, x_N]^T$ is an augmented covariate vector and $\mathbf{v} = [-\theta, w_1, \ldots, w_N]^T$ is an augmented weight vector. Each mixed dipole $(\mathbf{y}_i, \mathbf{y}_j)$, which should be divided, is associated with a function $\varphi_{ij}^m(\mathbf{v})$ being a sum of two functions with opposite signs ($\varphi_{ij}^m(\mathbf{v}) = \varphi_j^+(\mathbf{v}) + \varphi_i^-(\mathbf{v})$ or $\varphi_{ij}^m(\mathbf{v}) = \varphi_j^-(\mathbf{v}) + \varphi_i^+(\mathbf{v})$). For pure dipoles, which should stay undivided, we associate a function $\varphi_{ij}^p(\mathbf{v})$ ($\varphi_{ij}^p(\mathbf{v}) = \varphi_j^+(\mathbf{v}) + \varphi_i^+(\mathbf{v})$ or $\varphi_{ij}^c(\mathbf{v}) = \varphi_j^-(\mathbf{v}) + \varphi_i^-(\mathbf{v})$). A dipolar criterion function is a sum of penalty functions associated with each dipole:

$$\Psi_d(\mathbf{v}) = \sum_{(j,i) \in I_p} \alpha_{ij} \varphi_{ij}^p(\mathbf{v}) + \sum_{(j,i) \in I_m} \alpha_{ij} \varphi_{ij}^m(\mathbf{v}) \tag{3}$$

where α_{ij} determines relative importance (price) of the dipole $(\mathbf{y}_i, \mathbf{y}_j)$, I_p and I_m are the sets of pure and mixed dipoles, respectively. The parameters of neurons in a layer are obtained by sequential minimization (basis exchange algorithms [5]) of the dipolar criterion functions. The function is built from all the pure dipoles and those mixed dipoles which were not divided by previous neurons. The learning phase is finished when all the mixed dipoles are divided.

3 Optimization and neurons reduction phase

The neural network model created according to the procedure described in the previous paragraph may be over-fitted. To improve the generalization ability of the network the second phase of learning procedure - optimization - is proposed. The optimization (pruning) phase consists of two steps. The first step is aimed at distinguishing and enlargement of prototypes, and the other at reduction of redundant neurons.

The prototypes are the areas which contain the largest number of feature vectors $\mathbf{x}$. The number of prototypes is not fixed, but they should cover the assumed fraction of uncensored cases from the learning set (in the experiments 50%). Each prototype P is characterized by median survival time $Me(P)$ and 95% confidence interval for median survival time. The estimators of survival functions are calculated by using Kaplan-Meier method [8].

For each observation $o(\mathbf{x}, t, \sigma)$ from the learning set that is situated outside distinguished areas we need to choose the prototype to which the observation should belong. The prototype is chosen according to following rules:

1. determine the set of prototypes SP for which the observation o will not increase 95% confidence interval for median survival time
2. if $\sigma = 1$ choose the prototype $P_i \in SP$, for which

$$|t - Me(P_i)| = \min_{P_j \in SP} |t - Me(P_j)|$$

3. if $\sigma = 0$ choose the prototype $P_i \in SP$, for which $t - Me(P_i) < 0$ and

$$||\mathbf{x} - P_i|| = \min_{P_j \in SP} ||\mathbf{x} - P_j||$$

where $||\mathbf{x} - P_i||$ means the distance of the vector $\mathbf{x}$ to the prototype P_i

Location of the vectors $\mathbf{x}$ in the chosen prototypes is equivalent to minimizing a pice-wise linear criterion functions $Q_l(\mathbf{v}_l)$ $(l = 1, 2, \ldots, L)$ connected with all the neurons. As a results of minimization new prototypes are received. Whole procedure is repeated unless the global criterion function:

$$Q(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_L) = Q_1(\mathbf{v}_1) + Q_2(\mathbf{v}_2) + \ldots + Q_L(\mathbf{v}_L) \tag{4}$$

stops decreasing [4].

The other step of optimization phase causes the reduction of neurons which divide areas with similar survival times. As a similarity measure between two areas the log-rank test is used. The test verifies the hypothesis about no differences between survival times of two analyzed areas. The neuron is eliminated if in all the tests (number of tests is equal to the number of pairs of areas which will be joint after neuron reduction) the hypothesis was not rejected at 0.05 significance level.

After reduction of one neuron the whole optimization procedure is repeated from the beginning.

4 Experimental results

The analysis was conducted on the base of two data sets. The first data set is from the Veteran's Administration (VA) lung cancer study [8]. In this trial, male patients with advanced inoperable tumors were randomized to either standard (69 subjects) or test chemotherapy (68 subjects). Only 9 subjects from 137 are censored. Information on performance status at baseline (Karnofsky rating - KPS), disease duration in months, age in years at randomization, prior therapy (yes, no), and cell type (large, squamous, small, adeno), is available. The other dataset contains the information on 205 patients (148 censored cases) with malignant melanoma following radical operation. The data was collected at Odense University Hospital in Denmark by K.T. Drzewiecki [1]. Each patient is described by 4 features: sex, age, tumor thickness and ulceration.

The artificial neural network received for *VA lung cancer* data contains three formal neurons. Four areas were distinguished in the six dimensional feature space (see table 1). The best prediction is for patients belonging to the first area for which median survival time is equal to 200 days. For patients who belong to the area no. IV median survival time is equal to 19 days only. Survival functions calculated for all the areas are presented in figure 1a.

In *Malignant melanoma* dataset five homogeneous subgroups of patients were found. For the first three areas it is impossible to calculate median survival time because of the large fraction of censored cases in the data. Analyzing the survival

Table 1. Characteristics of the received patterns

Dataset	Pattern	Median (95% CI)	n (censored)
VA lung cancer	I	200 (162; 250)	25(3)
	II	117 (95; 132)	41(4)
	III	51 (35; 73)	40(0)
	IV	19 (13; 21)	31(2)
Malignant melanoma	I	-	50(44)
	II	-	19(16)
	III	-	69(48)
	IV	2567 (1228; ...)	16(9)
	V	793 (693; ...)	26(10)

functions (fig. 1b) we can see that the patterns no. I and III are very similar. They are represented by $[1, 1, 0]^T$ and $[1, 0, 1]^T$ network outputs respectively. The patters do not border directly on each other. It may suggest that there are two different areas with similar, long survival time in the feature space. In opposite to these two areas is subgroup no. V. The median survival time is equal to 793 days. The risk of failure for the patients belonging to this area is high.

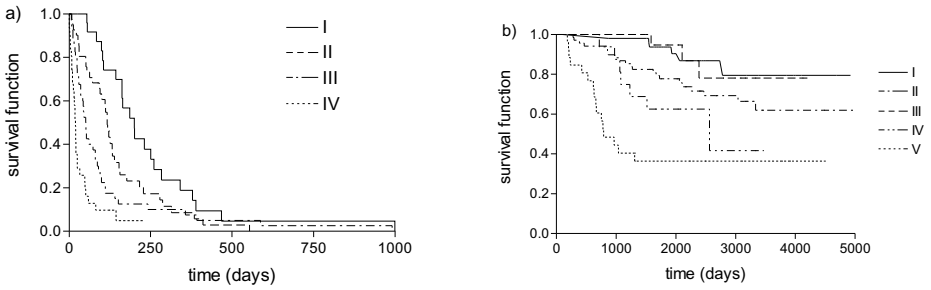

Fig. 1. Kaplan-Meier survival functions for distinguished areas: a) *VA lung cancer*; b) *Malignant melanoma*

5 Conclusions

The artificial neural network approach was proposed to identify areas in feature space which are homogeneous from the point of view of survival time. The main advantage of the method is lack of artificial division of survival time into disjoint intervals. The method based on the information received from the data identifies the subgroups of patients with similar survival experience. The algorithm both in the first and the other phase is able to deal with censored cases. Additional

two-step optimization phase allows improving the generalization ability of the network. Distinguished patters are defined by the rules possible to interpret also by people who are not closely related to the neural network methodology.

Acknowledgements This work was supported by the grant W/WI/1/02 from Białystok Technical University

References

1. Andersen P.K., Borgan O., Gill R. D., Statistical Models based on Counting Processes. Springer (1993)
2. Biganzoli E., Boracchi P., Mariani L., Marubini E., Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. Statistics in Medicine **17**(10) (1998) 1169–1186
3. Bobrowski L., Krętowska M., Krętowski M., Design of neural classifying networks by using dipolar criterions. Proc. of the Third Conference on Neural Networks and Their Applications, Kule, Poland (1997) 689–694
4. Bobrowski L., Krętowska M., Dipolar pruning of neural layers. Proc. of the 5th Conference on Neural Network and Soft Computing, Zakopane, Poland (2000) 174–179
5. Bobrowski L., Niemiro W., A method of synthesis of linear discriminant function in the case of nonseparability. Pattern Recognition **17** (1984) 205–210
6. Burke H.B., Goodman P.H., Rosen D.B., Henson D.E., Weinstein J.N., Harrell F.E., Marks J.R., Winchester D.P., Bostwik D.G., Artificial neural networks improve the accuracy of cancer survival prediction. Cancer **79** (1997) 857–862
7. De Laurentiis M., Ravdin, P.M., A technique for using neural network analysis to perform survival analysis of censored data. Cancer Letters **77** (1994) 127–138
8. Kalbfleisch J. D., Prentice R. L., The Statistical Analysis of Failure Time Data. John Wiley & Sons (1980)
9. Kappen H.J., Neijt J.P., Neural network analysis to predict treatment outcome. Annals of Oncology Suppl. **4** (1993) 31–34
10. Liestol K., Andersen P.K., Andersen U., Survival analysis and neural nets. Statistics in Medicine **13** (1994) 1189–1200
11. Mani D. R., Drew J.,Betz A., Datta, P., Statistics and data mining techniques for lifetime value modeling. Proc. of KDD (1999) 94–103
12. Ripley, B.D., Ripley, R.M., Neural networks as statistical methods in survival analysis. Artificial Neural Networks: Prospects for Medicine, Dybowski R., Grant V.(eds.), Landes Biosciences Publishers (1998)