

HIERARCHICAL NEURAL NETWORKS FOR SURVIVAL ANALYSIS

L. Ohno-Machado, M.G. Walker, and M.A. Musen

Section on Medical Informatics, Stanford University School of Medicine
MSOB X-215 Stanford CA 94305-5479, USA machado@camis.stanford.edu

ABSTRACT

Neural networks offer the potential of providing more accurate predictions of survival time than do traditional methods. Their use in medical applications has, however, been limited, especially when some data are censored or the frequency of events is low. To reduce the effect of these problems, we have developed a hierarchical architecture of neural networks that predicts survival in a stepwise manner. Predictions are made for the first time interval, then for the second interval, and so on. The system produces a survival estimate for patients at each interval, given relevant covariates, and is able to handle continuous and discrete variables, as well as censored data. We compared the hierarchical system of neural networks with a nonhierarchical system for a data set of 428 AIDS patients. The hierarchical neural-network model predicted survival more accurately than did the nonhierarchical model (although both had low sensitivity). The hierarchical model could also learn the same patterns in less than half the time that was required by the nonhierarchical model. These results suggest that the use of hierarchical systems is advantageous when censored data are present, the number of events is small, and time-dependent variables are necessary.

1. SURVIVAL ANALYSIS

Two functions of central interest in modelling survival data are the survivor function and the hazard function. The survivor function is defined as the probability that an individual survives at least up to a certain time. The hazard function represents the probability that an individual will die at a certain time, conditioned on his survival up to that time, and denotes the instantaneous death rate [1]. Survival analysis models, such as actuarial life-tables [2], product-limit estimators [3], proportional hazards [4], and fully parametric models [5] produce estimates for both the survivor function and the hazard function. Parametric methods of survival analysis require specification of a probability density function for estimating these functions. Nonparametric models do not require this specification, and are predominant in the biomedical literature. Researchers have recently applied new nonparametric models for survival analysis, such as regression trees [6] and neural networks [7], in medical data. Many concepts used commonly in classical models of survival analysis can also be employed in neural-network applications.

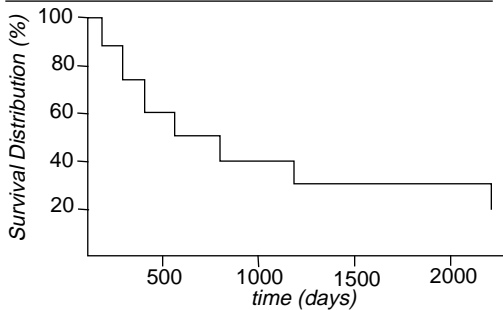
Actuarial life-tables and product-limit estimators of survivor functions are simple models that help researchers to summarize survival data. Both types of models involve the assumption that the survival of an individual at time t is conditioned on his survival at time $t-1$. The survivor function for actuarial life-tables and product-limit estimators is

$$S(t) = \prod \left(\frac{n_j - d_j}{n_j} \right)$$

where d_j is the number of deaths in interval j and n_j is the number of individuals at risk. In the actuarial method, n_j is the *average* number of individuals at risk. Actuarial life tables and product-limit estimators differ in the way that time intervals are built. The former model predefines intervals of equal duration and groups deaths in those intervals. The latter model builds one interval for each death, and therefore does not cause loss of information. Both models allow for censored data. Figure 1 displays a Kaplan–Meier survival curve for a cohort of AIDS patients. Neural network models, as we will demonstrate, also can estimate survivor functions in intervals when censored data are present. The Cox proportional hazards model is frequently used to study the importance of covariates for survival, and to produce survival prognoses. The Cox model is a multiple regression semi-parametric model that allows modeling of continuous covariates. It requires the assumption that there is a simplifying transformation of the initial data and that the hazards for the different individuals are proportional. A baseline hazard has to be estimated and hazards for individuals are multiples of the baseline

hazard. Although there are methods to test the validity of the assumptions [8], they are seldom used in practice. Hanson et al. described a cohort of HIV+ patients for whom the assumption of hazard proportionality does not hold [9]. For these cases, fully nonparametric models of survival that allow non-proportional hazards are needed. Actuarial life-tables and product-limit estimators could be used to produce survivor estimates for different strata, which should then be compared by the Mantel-Haenszel method [10], but this method cannot deal with continuous variables. Furthermore, stratifying data may result in few patients in each stratum, decreasing the statistical significance of the study.

Figure 1. Kaplan–Meier Survival Curve



Survival curve for the 428 AIDS patients of the ATHOS data set used in this study.

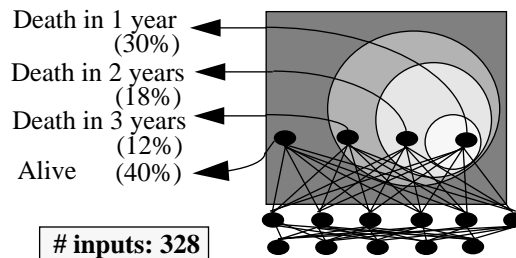
according to percentiles. The survival curves for the quartiles that generated similar prognoses were then plotted, and were compared with those of a Cox regression. There were no significant differences. The data probably respected the proportionality assumption, although this hypothesis was not tested explicitly. A bias was introduced by coding time as a covariate, and the authors had to balance the input data set to account for that bias. The importance of individual variables to the overall prediction of survival was not analyzed. The work of Ravdin et al. was one of the first studies to address the use of neural networks for survival analysis using real clinical data, producing accurate estimates for survival of breast cancer patients, and raising the important issue on how to deal with censored data in neural-network implementations for survival analysis. McGonigal et al. [15] have recently applied neural network models to assess the probability of survival for trauma patients, and the neural network model compared favorably with other systems. Their model also had a nonhierarchical implementation. Hierarchical implementations of neural networks have been developed for nonmedical problems. Curry et al. built a system of hierarchical neural networks to classify mass spectra [17], and showed that high classification accuracy was achieved. The hierarchical model was not compared to a nonhierarchical one. In this paper, we compare the performance of these two types of models for survival prediction of AIDS patients.

2. HIERARCHICAL NEURAL NETS FOR SURVIVAL ANALYSIS

We have shown previously that the decomposition of a classification task in small subcomponents facilitates learning in neural network systems, especially when there are hierarchies of solutions and the frequency (prior probability) of certain categories is low [18,19]. A similar approach can be used in survival analysis, where the frequency of events in each interval may be low compared to the overall number of individuals. To test whether differences in pattern frequency would affect the behavior of our network in terms of the time it took for the networks to identify rare events, we designed the following experiment. We developed three systems of neural networks to perform survival analysis. In a nonhierarchical system, output nodes correspond to the probability that an individual will die during the

The most common applications of neural networks in clinical medicine have been for diagnosis of diseases. A few reports on the use of neural networks for other kinds of classification tasks have been published. These systems determine prognosis after cardiopulmonary resuscitation [11], strategies for weaning from respiratory support [12], tumor stage in oncology [7,13], graft outcome after liver transplantation [14], and prognosis in trauma [15]. Ravdin et al. [7,16] studied the prognosis of breast-cancer patients using neural network models. In their models, time was entered as a predictor, and each patient had as many entries in the model as the number of intervals during which she was alive. The intervals were derived from Kaplan–Meier estimates,

Figure 2. Nonhierarchical Neural Network

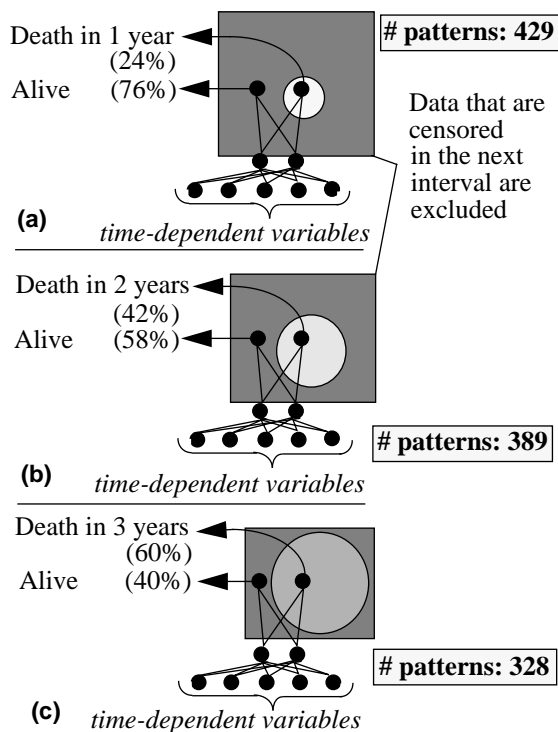


This neural network predicts which patients died within one, two, and three years of follow-up. The square represents the number of patients who were alive. The smaller circle represents the patients who died in the first year, the intermediate circle represents the ones who died in the second year, and so on. Censored data were not used.

first interval, second interval, and so on. Outputs are mutually exclusive. The training data correspond to individuals who have been followed for as many intervals as there are output categories (e.g., “alive,” “death in one year,” “death in two years”). In a nonhierarchical model, Type I censored data (data from patients followed for a short interval) are not used, since the status of the patient during the last intervals is not known. Figure 2 displays a nonhierarchical model.

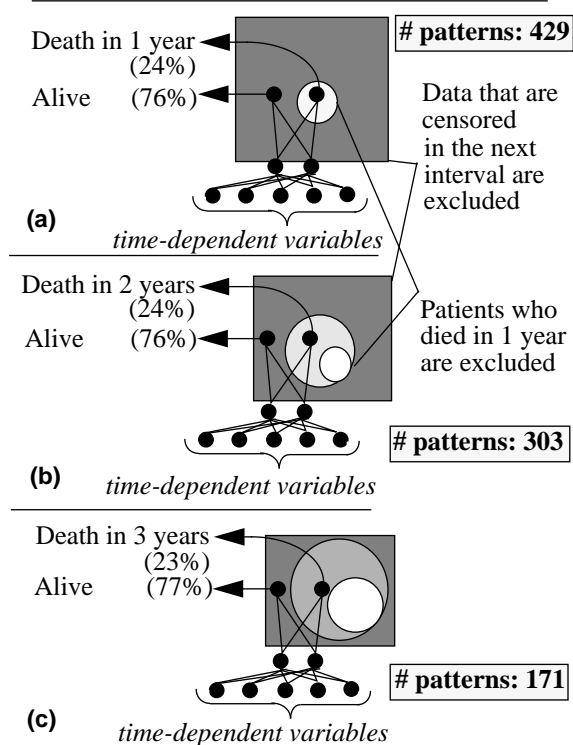
Two other systems predict prognoses hierarchically. In these systems, each of the component networks corresponds to an interval in time. One of the systems predicts cumulative survival (that is, the system tries to predict if a certain patient will be dead after an interval of time), and the other predicts the conditional survival (that is, the probability that a certain individual who has survived up to a certain time will survive up to the next interval). In the former case, even though the event frequency (i.e., the number of deaths) may be low for the first intervals, the frequency in the late intervals is increased due to accumulation. In both cases, even though the number of events is low compared to the total number of individuals, we consider only the number of events relative to the number of individuals at risk during each interval, so the frequency of events is proportionately increased in each of the component networks. Censored data are allowed in the hierarchical systems. Figure 3 displays a hierarchical system to predict absolute, cumulative survival. Figure 4 displays a hierarchical system to predict instantaneous, conditional survival.

Figure 3. Hierarchical Neural Network for Predicting Absolute Survival



This system of neural networks predicts survival in intervals of (a) 1 year, (b) 2 years, and (c) 3 years. Squares correspond to the patients who were alive, and the circles correspond to patients who died in the interval. Data censored in the second year are used in the first network (429 patients). Data censored in the third year are used in the second network (389 patients). Data used in the third network are the same as those used in the nonhierarchical model in Figure 2.

Figure 4. Hierarchical Neural Network for Predicting Conditional Survival



This system of neural networks predicts the death of a patient conditioned on his survival up to that time. Squares correspond to patients who were alive and circles correspond to patients who died during the interval. Figure 4(a) represents the same patients as 3(a), representing predictions for the first year. In (b), the smaller circumference represents the patients who died in the first interval and therefore do not compose the input patterns in the second year. Patients who were alive in the first year and died in the second year are represented by the solid part of the large circle. In (c), patients who died during the second interval are excluded. Censored data are treated as the model shown in Figure 3.

The training data correspond to individuals who have been followed for at least the same number of days as that of the interval of prediction. Data from censored individuals are used, but the data are presented to the network only up to the point where the data are confirmed. Therefore, the neural networks that are higher in the hierarchy (those predicting the event in the first intervals) receive more input patterns than do the ones that are lower in the hierarchy. Each of the component networks may receive different values for time-dependent variables for the same individual. Our neural-network systems can model nonlinear relationships between continuous covariates and outcomes. The hierarchical neural-network systems are able to use Type I censored data and can potentially deal with time-dependent variables. We studied the ability of our models to fit the data, and to generalize to new cases. The following compares these three systems for a database of HIV positive patients.

3. NEURAL NETWORKS FOR SURVIVAL ANALYSIS IN HIV INFECTION

The ATHOS database is a prospective, longitudinal, primary data set of HIV+ and at-risk subjects, collected from 10 clinics in California, under the direction of Dr. James Fries [20]. Data describing a subset of 428 patients from the ATHOS collection were used in the present experiment. All patients in the ATHOS data set for whom the values of the independent variables were known, and who has been followed for more than one year were selected. The Kaplan–Meier survival curve for this subset is shown in Figure 4. We predicted death in one, two, and three years for this subset of patients using the hierarchical model, and compared its performance with that of a standard nonhierarchical neural network model. We divided the data set into a training set (entries with odd numbers) and a test set (entries with even numbers). Parameters were estimated using the training set and evaluation used the test set. The independent variables included age, gender, ethnic origin, risk behavior, educational level, AIDS-defining diagnoses, and duration of AIDS. The dependent variable was death caused by AIDS-related conditions in each time interval. No time-dependent variables were used in this experiment. The frequencies for each event in our three models are shown in Figures 2, 3 and 4.

Each network of the hierarchical system was composed of twenty-seven input nodes, four hidden nodes, and two output nodes. Input values for age and for duration of AIDS were continuous, whereas input values for gender, ethnic origin, risk behavior, educational level, AIDS-defining diagnoses, and outputs for “alive” or “death” at each interval were coded by dummy variables. The nonhierarchical model had twenty-seven input nodes, ten hidden nodes, and two output nodes, so that it had approximately the same number of weights as the hierarchical system. Both systems used the backpropagation algorithm [21] for estimating the weights. The learning rate for both systems was 0.01, no momentum term was added, and the cross-entropy error function was used. Output values corresponded to the probability of a given category (e.g., “dead in one year,” “alive”), and always summed to 1.0. We studied the ability of our networks to learn the death patterns using the training set, and their ability to generalize using the test sets.

The nonhierarchical model that predicted conditional survival took approximately 32,000 epochs to recognize at least half of the patients who died in the third year (the least frequent output category, representing only 12% of the patterns). The hierarchical models took approximately 16,000 epochs. Since the time required per epoch in the hierarchical system is smaller than that of the nonhierarchical system, this difference showed that the hierarchical models could learn the same patterns in less than half of the time that required by the nonhierarchical model. We believe that this difference is due primarily to the different frequencies involved in the input data for each system. This hierarchical system could correctly identify nine patients out of thirty who died in the third year, whereas the nonhierarchical model could recognize only five. The specificity, or true-negative rate, of the hierarchical model at this level was superior (0.95 and 0.87). The hierarchical model was also able to recognize more patients who died in the second year than could the nonhierarchical model (12 and 7, respectively, corresponding to true-negative rates of 0.75 and 0.79). The performance was approximately the same as that of the nonhierarchical model for detecting patients who died during the first year of disease (17 patients recognized for the hierarchical system and 20 for the nonhierarchical system, with true-negative rates of 0.81 and 0.71, respectively). In the latter case, the pattern frequencies were approximately the same in both systems, so no improvement was indeed expected.

The nonhierarchical model that predicted absolute (or “unconditional”) survival was even better for detecting deaths in the second and third years than the nonhierarchical model, because for these intervals the cumulated percentage of deaths (see Figure 4) made this category of patterns more easily identifiable. These preliminary results support our hypothesis that neural-network systems are able to recognize infrequent patterns after a number of training cycles that is inversely proportional to the frequency of those patterns. The

overall sensitivities and specificities of the hierarchical models, although superior to those of the nonhierarchical models, were small. The limited use of data and the small number of patients may have caused the poor prognosis prediction in this preliminary experiment. The hierarchical systems of neural networks, by increasing the relative frequency of the outcome being sought, can greatly increase the speed of learning, whereas they do not hinder accuracy performance, when compared to nonhierarchical neural-network systems.

4. CONCLUSION

We compared three neural network models for survival analysis. For a cohort of 428 AIDS patients, the hierarchical neural-network models for survival analysis could learn infrequent patterns faster than could a nonhierarchical model. The hierarchical models also provided better accuracy in predicting death for this cohort.

5. REFERENCES

1. Collet D. *Modelling Survival Data in Medical Research*. Chapman and Hall, London, 1994.
2. Cutler SJ; Ederer E. Maximum utilization of the life-table in analysis of survival. *Journal of Chronic Diseases*, 1958, 6:699-712.
3. Kaplan EL; Meier P. Nonparametric estimation from incomplete observations. *American Statistical Association Journal*, 1958 June, 457-81.
4. Cox DR. *Analysis of Survival Data*. London, Chapman and Hall, 1984.
5. Lee ET. *Statistical Methods for Survival Data Analysis*. John Wiley and Sons, 1992.
6. Segal MR; Bloch DA. A comparison of estimated proportional hazards models and regression trees. *Statistics in Medicine*, 1989, 8:539-50.
7. Ravdin PM; Clark GM; Hilsenbeck SG; Owens MA; Vendely P; Pandian MR; McGuire WL. A demonstration that breast cancer recurrence can be predicted by neural network analysis. *Breast Cancer Research and Treatment*, 1992, 21(1):47-53.
8. Chale JJ; Quantin C; Mosseri V; Asselain B; Moreau T; Dussere L. Testing the proportional hazards hypothesis on a tonsillar carcinoma data set: A Comparison of Methods. *Proceedings of the Seventh World Congress on Medical Informatics*, Geneva, 1992, 890-4.
9. Hanson DL; Horsburgh CR Jr; Fann SA; Havlik JA; Thompson SE 3d. Survival prognosis of HIV-infected patients. *Journal of Acquired Immune Deficiency Syndromes*, 1993 Jun, 6(6):624-9.
10. Mantel N; Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 1959, 22, 719-48.
11. Ebell MH. Artificial neural networks for predicting failure to survive following in-hospital cardiopulmonary resuscitation. *Journal of Family Practice*, 1993 Mar, 36(3):297-303.
12. Ashutosh K; Lee H; Mohan CK; Ranka S; Mehrotra K; Alexander C. Prediction criteria for successful weaning from respiratory support: statistical and connectionist analyses. *Critical Care Medicine*, 1992 Sep, 20(9):1295-301.
13. Burke HB. Artificial neural networks for cancer research: outcome prediction. *Seminars in Surgical Oncology*, 1994, 10:73-9.
14. Doyle HR; Dvorchik I; Mitchell S; Marino IR; Ebert FH; McMichael J; Fung JJ. Predicting outcomes after liver transplantation. *Annals of Surgery*, 1994, 219(4):408-15.
15. McGonigal MD; Cole J; Schwab CW; Kauder DR; Rotondo MF; Angood PB. A new approach to probability of survival scoring for trauma quality assurance. *Journal of Trauma*, 1993 Jun, 34(6):863-8.
16. Ravdin PM; Clark GM. A practical application of neural network analysis for predicting outcome of individual breast cancer patients. *Breast Cancer Research and Treatment*, 1992, 22(3):285-93.
17. Curry B; Rumelhart DE. MSnet: A neural network that classifies mass spectra. *Tetrahedron Computer Methodology*, 1990, 3:213-37.
18. Ohno-Machado L; Musen MA. *Hierarchical Neural Networks for Partial Diagnosis in Medicine*. Proceedings of the World Congress on Neural Networks, San Diego, 1994, vol.I:291-6.
19. Ohno-Machado L. Identification of Low Frequency Patterns in BackPropagation Neural Networks. Technical Report KSL-94-29, Knowledge Systems Laboratory, 1994.
20. Fries J. *An HIV Data Resource for Systematic Health Policy Study*. Grant Proposal for the AHCPR, 1992.
21. Rumelhart DE; Hinton GE; Williams RJ. Learning internal representation by error propagation. In Rumelhart, D.E., and McClelland, J.L. (eds) *Parallel Distributed Processing*. MIT Press, Cambridge, 1986.

Acknowledgments

The authors thank Dr. D. Lubeck, Dr. J. Fries, and Mr. P. O'Driscoll for the use of the ATHOS data set, and Dr. D. Rumelhart for his advice on neural network implementations. This work has been funded by the Conselho Nacional de Pesquisa (CNPq), Brazilian Ministry of Education. Computing facilities were provided by the CAMIS resource, through grant LM05305 from the National Library of Medicine.