Application of Random Forests and Deep Neural Networks to Suicide Death Data

STAT-6494 Project Report

Wenjie Wang[*]

29 April 2018

Abstract

The classical survival models, such as Cox proportional hazard model, often require extensive efforts on variable selection or prior medical information to model interaction between patients' covariates and treatment covariates. While nonlinear models, such as neural networks and random forests, are able to model high-order interaction terms. It is of interest to apply these machine learning methods to survival data and compare their performance with classical statistical models.

Keywords: Cox Model, Machine Learning, Suicide Prevention

1 Introduction

For survival data, medical researchers' interests often lie in discovery of significant treatment effects and important diagnosis covariates of patients. The classical survival models, such as Cox proportional hazard model, assume risk function in a simple linear form of covariates, which can be too simplistic to capture the underlying relationship between response and covariates. In addition, they often require extensive efforts on variable selection or prior medical information to model interaction between patients' covariates and treatment covariates. While nonlinear models, such as neural networks and random forests, are able to model high-order interaction terms. It is of interest to apply these machine learning methods to survival data and compare their performance with classical statistical models. It would be even more interesting to discover nonlinear relationship by machine learning methods and build a statistical model for better interpretation and capability for statistical inferences.

2 Random Forests for Survival Data

Compared with linear models, tree models are able to incorporate complex interaction between covariates more naturally. Breiman (1996) proposed a bootstrap aggregation method named Bagging that aggregates predictors based on bootstrap samples to improve prediction accuracy or reduce the variance of an estimated prediction function, which is an early example of ensemble methods. Breiman (2001) proposed random forests (RF), an ensemble tree method that introduces randomization to the base learning process. The randomness lies in two folds: each tree is trained against a bootstrap sample; within each tree, the candidate variables are randomly selected at each node, among which the best split is found to maximize the difference of daughter nodes. Compared with Bagging, the additional randomness introduced to random forests results in predictors that

[*]wenjie.2.wang@uconn.edu; Ph.D. student at Department of Statistics, University of Connecticut.

ensembles less correlated trees and is showed to further improve prediction accuracy. Ishwaran et al. (2008) extended RF method to random survival forests (RSF) method for analysis of right-censored survival data. The algorithm can be summarized as follows:

1. draw B bootstrap samples from the origin data
2. grow a survival tree with randomly selecting p candidate covariates at each node and maximizing survival difference between daughter nodes
3. each terminal node: no less than $d_0 > 0$ unique deaths
4. compute cumulative hazard function (CHF) by Nelson-Aalen estimator for each tree and average to obtain the ensemble CHF
5. compute prediction error using out-of-bag (OOB) data

The difference between original RF model and RSF model is mainly on the splitting rules and the ensemble function. Ishwaran et al. (2008) considered four different splitting rules for RSF:

1. the log-rank splitting rule (Segal, 1988) that has been shown to be robust in both proportional and non-proportional hazard settings (LeBlanc and Crowley, 1993);
2. the conservation-of-events splitting rule that splits nodes by finding daughters satisfying the conservation-of-events principle introduced in (Naftel et al., 1985), which asserts that the sum of the estimated CHF must equal the total number of deaths;
3. the log-rank score rule splitting nodes by a standardized log-rank statistic (Hothorn and Lausen, 2003);
4. the random log-rank splitting rule using a random split with maximum log-rank statistic among the candidate variables for each node.

The random survival forests model does not assume proportional hazards and thus may be more attractive than regular Cox Proportional hazard model (Cox, 1972). In addition, it naturally takes into account interaction effects between covariates and providers insights on the importance of covariates based on the tree structures. Variable selection can be performed based on the variable importance (VIMP) measure. Similar to the regular RF, the VIMP can be computed for RSF by conceptually dropping OOB samples down the in-bag survival tree. The VIMP for variable x is the prediction error for the original ensemble subtracted from the prediction error for the new ensemble obtained using randomizing x assignments (Ishwaran and Kogalur, 2007). Ishwaran et al. (2010) further proposed an algorithm named "RSF-Variable Hunting" for high-dimensional variable selection for survival data based on the idea of minimal sub-tree (Ishwaran et al., 2007).

3 Deep Neural Networks for Survival Data

The regular Cox proportional hazards model has a linear relative risk function $r(\boldsymbol{x}, \boldsymbol{\beta}) = \boldsymbol{\beta}^\top \boldsymbol{x}$. In many applications, it is hard to assume a linear proportional hazards condition and thus high-level interaction terms are required. However, as the number of covariates and interactions increases, it becomes prohibitively expensive.

Katzman et al. (2016) proposed a Cox proportional hazards deep neural network method called DeepSurv for personalized treatment recommendations. DeepSurv is a multi-layer perceptron that predicts a patient's risk of death. The output of the network is a single node estimating the relative risk function $\hat{r}_\theta$ by the weights of the network θ. The loss function is set to be the negative log partial likelihood (Cox, 1975). It allows more than one layer and outputs a single node estimating the relative risk function $\hat{r}_\theta$ by the weights of the network θ. Quite a few modern techniques

are applied in the model fitting, such as weight decay (L2-norm) regularization, Rectified Linear Units (ReLU) with batch normalization (Ioffe and Szegedy, 2015), dropout (Srivastava et al., 2014), gradient descent optimization algorithms including stochastic gradient descent and adaptive moment estimation named Adam (Kingma and Ba, 2014), Nesterov momentum (Nesterov, 2013), gradient clipping (Pascanu et al., 2012), learning rate scheduling (Senior et al., 2013), etc.

4 Connecticut Suicide Death Data

Suicide is a serious public health problem in the US. Death by suicide is increasing among all age groups in the US, with a 24% increase in suicide rates observed from 1999 to 2014. There is a strong likelihood that suicide attempters would make additional attempts after the initial suicide attempt (Suominen et al., 2004), and suicide attempt is a strong predictor of suicidal death (Bostwick et al., 2015). The subjects in the suicide death data were patients in the State of Connecticut who have been hospitalized for suicide attempt or intentional self-injury during fiscal year 2005 to 2012 (from October 1, 2004 to September 30, 2012). Data from diagnosis were available from the Connecticut Hospital Inpatient Discharge Data (HIDD). Deaths by suicide were determined from the Office of the Connecticut Medical Examiner (OCME). We are interested in the time since hospitalization due to suicide attempts to suicidal death of those patients. A total of 22,221 patients were followed up until September 30, 2012. Among them, 16,208 (73%) were white (9,108 female and 7,100 male) and 6,013 (27%) were non-white (3,220 female and 2,793 male). The number of event (suicidal death) was only 606 and thus the censoring rate was about 97.3%. The Kaplan-Meier survival curve is given in Figure 1.

The HIDD data contained a large number of records on the characteristics of patients and their previous hospital admissions. One of the research interest was to identify important diagnostic categories associated with patient death. The diagnostics were recorded as ICD-9 diagnosis codes, or more formally ICD-9-CM (International Classification of Diseases, 9th Revision, Clinical Modification). We grouped the ICD-9 codes by their three leading characters that define the major diagnosis categories. Suicide attempts were identified by both ICD-9 external cause of injury codes and other ICD-9 code combinations indicative of suicidal behavior (Patrick et al., 2010; Chen and Aseltine, 2017). Other ICD-9 codes during the inpatient hospitalization fell into 167 major diagnosis categories, which led to 167 indicator variables.

4.1 Results

We randomly split the 70% of suicide death data to a training set and the remaining 30% to a test set 200 times. In each random split, we fitted the RSF model and the DeepSurv model on the training set and measured the predicting power of these two methods over the testing set through Harrell's c-statistic (Harrell et al., 1996), an extension of the area under the receiver-operator characteristics curve for censoring data. This statistic is an estimate of the probability of concordance between the order of risk scores and survival outcomes. For the RSF, the average c-statistic of the training set and test set was 0.68 and 0.72, respectively. While for the DeepSurv model, the average c-statistic of the training set and test set was 0.56 and 0.58, respectively. A side-by-side boxplot of the training and testing c-statistic from 200 random splits is given in Figure 2, from which we find that RSF model provides a better prediction performance than the DeepSurv model on the suicide dataset.

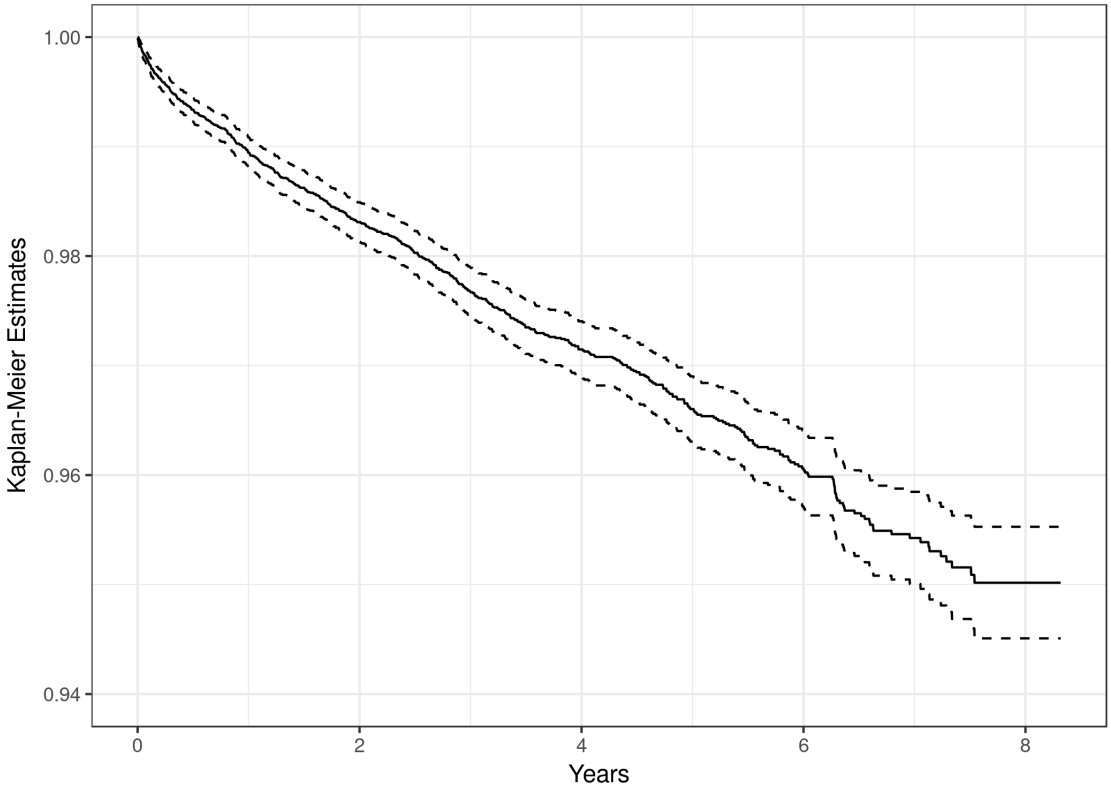

Figure 1: The Kaplan-Meier survival curve of the CT suicide data with 95% confidence band.

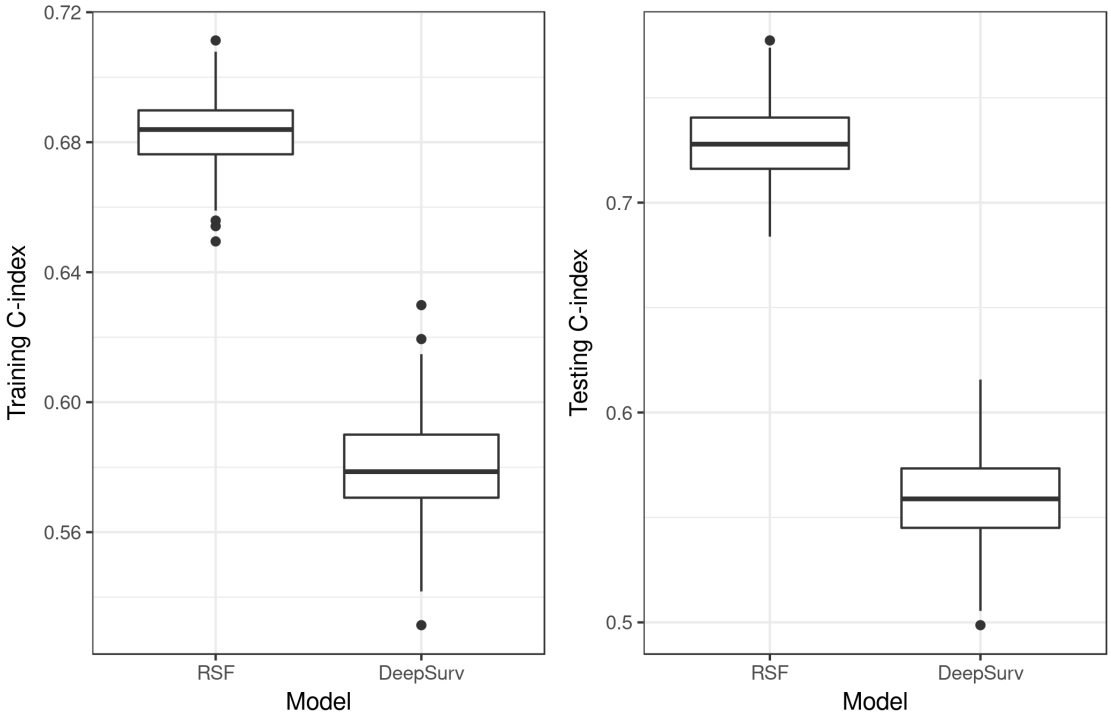

Figure 2: A side-by-side boxplot of the training and testing c-statistic from 200 random splits.

In addition, we performed variable selection from RSF model by the minimal depth method (Ishwaran et al., 2010). For the training data from each random split, we obtained the averaged number of maximal sub-trees (normalized by the size of a tree) for each variable. Then we computed the sum of these average numbers over all random splits. Length of stay, age, gender (male), and race (white) are all among the top 30 variables having the largest sum. The top ICD-9 diagnosis codes having the largest sum include E84 (air and space transport accidents), 305 (nondependent abuse of drugs), 296 (episodic mood disorders), 969 (poisoning by psychotropic agents), 965 (poisoning by analgesics antipyretics and antirheumatics), 311 (depressive disorder, not elsewhere classified), 780 (general symptoms), 300 (anxiety, dissociative and somatoform disorders), 304 (drug dependence), 301 (personality disorders), E85 (accidental poisoning by drugs, medicinal substances, and biologicals), 276 (disorders of fluid electrolyte and acid-base balance), 303 (alcohol dependence syndrome), 401 (essential hypertension), 309 (adjustment reaction), 881 (open wound of elbow forearm and wrist), v62 (other psychosocial circumstances), E98 (injury Undetermined Whether Accidentally Or Purposely Inflicted), 518 (other diseases of lung), 493 (asthma), and 292 (drug-induced mental disorders), etc.

5 Summary and Discussion

In this course project, we applied random survival forests and deep neural network model to the suicide death data and obtained some hands-on experience of the "state-of-the-art" models. Compared with deep neural network model, RSF is closer to the so-called "off-the-shelf" procedure for data mining since it provides insights for variable selection and does not require heavy tuning on the hyper-parameters.

Both methods are quite computationally intensive. We applied parallel computing on the data level for fitting the 200 random splits of the suicide data. Fitting RSF models with only 1,000 trees over 200 random splits using six cores (Intel i7, up to 3.8 GHz) took about 30 hours, while Fitting DeepSurv models with only two hidden layers using four cores took about 50 hours. Due to the limitation of available computing resource, we did not perform systematic tuning procedure, such as cross-validation and random hyper-parameter optimization search (Bergstra and Bengio, 2012) for hyper-parameters in deep neural network model, which was the probably the reason why we did not obtain comparable prediction performance from the DeepSurv model.

We also performed a simulation study for investigating the prediction performance of the RSF model and the DeepSurv model over survival data with a fraction of cure group. However, the results are not included and not discussed in this report. All the source code is available at the project repository on GitHub: https://github.com/statds/final-project-wenjie_wang. The RSF model has an existing implementation in R and the DeepSurv model has existing implementation in Python. However, setting up a working and probably reproducible computing environment for both models is not trivial. Therefore, we set up a Dockerfile and built a docker image for this project for providing a readied and reproducible (hopefully) computing environment for people who are interested in this project. The docker image is available at Docker Hub: https://hub.docker.com/r/wenjie2wang/statds-spring2018/.

Acknowledgment

The author would like to thank Professor Jun Yan, Professor Elizabeth Schifano, and Professor Kun Chen for providing this interesting and inspiring course. All comments and suggestions received (from classmates) were/would be also appreciated.

Reference

Bergstra, J. and Bengio, Y. (2012), "Random Search for Hyper-Parameter Optimization," *Journal of Machine Learning Research*, 13, 281–305.

Bostwick, M. J., Pabbati, C., Geske, J. R., and McKean, A. J. (2015), "Suicide Attempt as a Risk Factor for Completed Suicide: Even More Lethal Than We Knew," *The American Journal of Psychiatry*, 173, 1094–1100.

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24, 123–140.

— (2001), "Random Forests," *Machine Learning*, 45, 5–32.

Chen, K. and Aseltine, R. (2017), "Using Hospitalization and Mortality Data to Target Suicide Prevention Activities: A Demonstration from Connecticut," *Journal of Adolescent Health*, 61, 192–197.

Cox, D. R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, 34, 187–220.

— (1975), "Partial Likelihood," *Biometrika*, 62, 269–276.

Harrell, F. E., Lee, K. L., and Mark, D. B. (1996), "Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors," *Statistics in medicine*, 15, 361–387.

Hothorn, T. and Lausen, B. (2003), "On the Exact Distribution of Maximally Selected Rank Statistics," *Computational Statistics & Data Analysis*, 43, 121–137.

Ioffe, S. and Szegedy, C. (2015), "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *International conference on machine learning*, pp. 448–456.

Ishwaran, H. and Kogalur, U. B. (2007), "Random survival forests for R," *R News*, 2, 25–31.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008), "Random Survival Forests," *The annals of applied statistics*, 841–860.

Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J., and Lauer, M. S. (2010), "High-Dimensional Variable Selection for Survival Data," *Journal of the American Statistical Association*, 105, 205–217.

Ishwaran, H. et al. (2007), "Variable Importance in Binary Regression Trees and Forests," *Electronic Journal of Statistics*, 1, 519–537.

Katzman, J., Shaham, U., Bates, J., Cloninger, A., Jiang, T., and Kluger, Y. (2016), "DeepSurv: Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network," *ArXiv e-prints*.

Kingma, D. P. and Ba, J. (2014), "Adam: A Method for Stochastic Optimization," .

LeBlanc, M. and Crowley, J. (1993), "Survival Trees by Goodness of Split," *Journal of the American Statistical Association*, 88, 457–467.

Naftel, D., Blackstone, E., and Turner, M. (1985), "Conservation of events," Unpublished notes.

Nesterov, Y. (2013), "Gradient Methods for Minimizing Composite Functions," *Mathematical Programming*, 140, 125–161.

Pascanu, R., Mikolov, T., and Bengio, Y. (2012), "Understanding the Exploding Gradient Problem," *CoRR*, abs/1211.5063.

Patrick, A. R., Miller, M., Barber, C. W., Wang, P. S., Canning, C. F., and Schneeweiss, S. (2010), "Identification of Hospitalizations for Intentional Self-Harm When E-codes are Incompletely Recorded," *Pharmacoepidemiology and Drug Safety*, 19, 1263–1275.

Segal, M. R. (1988), "Regression Trees for Censored Data," *Biometrics*, 35–47.

Senior, A., Heigold, G., Yang, K., et al. (2013), "An Empirical Study of Learning Rates in Deep Neural Networks for Speech Recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, IEEE, pp. 6724–6728.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014), "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *The Journal of Machine Learning Research*, 15, 1929–1958.

Suominen, K., Isometsä, E., Suokas, J., Haukka, J., Achte, K., and Lönnqvist, J. (2004), "Completed Suicide After a Suicide Attempt: A 37-Year Follow-Up Study," *American Journal of Psychiatry*, 161, 562–563.