Random Survival Forests

Jeremy M. G. Taylor, PhD

In the article by Chen et al,[1] the authors used Random Survival Forests (RSF) as part of their approach for analyzing the data. In this note, we will explain RSF in a nontechnical way; precise details of the RSF method are described in the article by Ishwaran et al.[2] RSF are an adaptation of Random Forests (RF)[3] designed to be used for survival data. Software to run RSF is described in the article by Ishwaran and Kogalur.[4] RSF differs from RF in that the response data are a survival time, which may be censored. RF are a group of methods that are developed from data in which there is an outcome or response variable and also a potentially large number of predictors or explanatory factors. The dataset would typically include data from a large number of subjects. In the article by Chen et al, the explanatory factors are all the gene expression measurements, after some initial selection of genes, age, gender, and stage, and the response variable is survival time. The model is designed to be used for prediction purposes. Specifically for a new subject, who has all the explanatory factors measured, the model gives a prediction of the response.

The RF method has a number of appealing features. One major feature is that it can easily handle datasets with many more variables than subjects. It does not delete or select any of the variables, it allows them all to influence the prediction if the training data suggest they should. Another attractive feature of RF is that they do not impose a restrictive structure on how the variables should be combined, for example, it is not a weighted sum of the gene expression values. If the relationship between the predictor variables and the response variable is complex with nonlinear patterns and interactions then RF are capable of incorporating this.

RF is an ensemble tree method, i.e., the Forest consists of many trees. The final prediction is a combination of the predictions from each tree. It is well known that methods that combine predictions from separate methods can substantially improve prediction performance. It has also been shown that injecting some controlled variation or randomness into the construction of each of the separate trees can improve performance. Hence the name Random Forests. Each tree in the Forest is slightly different, and each tree by itself gives a prediction. The final prediction from the Forest is the average of the predictions from each tree. Each decision tree is a series of simple questions with yes/no answers, where examples of the questions could be "Is gene-78 $>$ 2.7?" or "Is gene-953 $>$ 5.3?" When an observation, consisting of a set of gene expression measurements, is "dropped down the tree," the questions are asked sequentially about that observation, the answers to the questions determine the path through the tree until all questions have been asked for that path. All observations that follow the same path will be similar and will be given a predicted value for the response. How each tree is constructed is a complex procedure, but the general idea is to choose the questions (e.g., gene 78 and gene 953) and the cut-points (e.g., 2.7 and 5.3) so the group who answer "Yes" has a very different outcome variable value from the group who answer "No."

The RSF produces a prediction for each observation that is dropped down the tree, this prediction we refer to as the mortality risk index (MRI). A higher MRI means this person is at higher risk based on their gene expression values. The actual definition of the MRI is the expected number of deaths that would be seen if all observations in the sample had that particular set of gene expression values. The MRI is thus a single number that can be used to put people into categories of low, medium, or high, for example.

Because the inner workings of a RF are not transparent, trying to understand which of the input variables are important in determining the predictions is more challenging. However, a measure of how important a variable is can be provided by a RF, and it is called the variable importance measure (VIMP). It measures how worse would the prediction be if that variable were not available. In the article by Chen et al, not surprisingly stage had the highest VIMP values and age was one of the highest. We also used the VIMP in the article by Chen et al to help in eliminating genes. Specifically, we added a set of irrelevant random variables as potential predictors. The VIMP for those random variables were very low as we expected. This helps us set a threshold for the VIMP, as we can select real predictors whose VIMP is greater than the VIMP of the random variables.

Another feature that is provided by RSF is the prediction error. For other prediction models, this is called the C-index and is related to the area under the receiver operating characteristic curve. The prediction error is the fraction of times that for a pair of subjects the person who was predicted to live longer actually died sooner, i.e., the prediction ranked these two people incorrectly. Thus, a small prediction error rate is good, and error rates of less than 25% would be

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan.
Disclosure: The author declares no conflicts of interest.
Address for correspondence: Jeremy M. G. Taylor, PhD, Department of Biostatistics, University of Michigan, Ann Arbor, MI, 48109. E-mail: jmgt@umich.edu

ISSN: 1556-0864/11/0612-1974

desirable. Prediction error rates of 50% or higher are useless because they are no better than tossing a coin. In the article by Chen et al, the prediction error rates were in the range of 25 to 40%, which indicates some promise for the predictor, but far from ideal and indicative of the complexity of the predicting patient survival outcome.

When the goal is prediction, RSF are an attractive method to build a model. They are particularly useful in situations where there are a large number of predictors and the relationship between the response and the predictors may be complicated. As with all models that use a large number of predictors, datasets with a large number of subjects are required if one expects to develop a reliable predictor.

REFERENCES

1. Chen G, Kim S, Taylor JMG, et al. Development and validation of a qRT-PCR classifier for lung cancer prognosis. *J Thorac Oncol* 2011;6:1481–1487.
2. Ishwaran H, Kogalur UB, Blackstone EH, et al. Random survival forests. *Ann Appl Stat* 2008;2:841–860.
3. Breiman L. Random forests. Machine learning, 2001;45:5–32.
4. Ishwaran H, Kogalur UB. Random survival forests for R. Rnews. 2007;7:25–31.

