Applying novel tree-based frameworks to big data for classification of heart failure patients and prediction of clinical responses

Yan Zhang  
*Yale University, yan.zhang.yz464@yale.edu*

Nicholas Downing  
*Yale University*

Emily Bucholz  
*Yale University*

Suganithi Balasubramanian  
*Yale University*

Shu-Xia Li  
*Yale University*

*See next page for additional authors*

Follow this and additional works at: [http://elischolar.library.yale.edu/dayofdata](http://elischolar.library.yale.edu/dayofdata)

Part of the [Bioinformatics Commons](http://elischolar.library.yale.edu/dayofdata), [Cardiovascular Diseases Commons](http://elischolar.library.yale.edu/dayofdata), [Health Information Technology Commons](http://elischolar.library.yale.edu/dayofdata), and the [Public Health Commons](http://elischolar.library.yale.edu/dayofdata)

---

[http://elischolar.library.yale.edu/dayofdata/2014/Posters/13](http://elischolar.library.yale.edu/dayofdata/2014/Posters/13)

---

This Event is brought to you for free and open access by EliScholar – A Digital Platform for Scholarly Publishing at Yale. It has been accepted for inclusion in Yale Day of Data by an authorized administrator of EliScholar – A Digital Platform for Scholarly Publishing at Yale. For more information, please contact elischolar@yale.edu.
Use Case Background

Over 5 million Americans suffer from heart failure (HF), a condition with a 5-year survival that eclipses all cancers apart from that of lung cancer. Conventional understanding of HF is simplistic: it is viewed as a single syndrome, despite real heterogeneity. A novel approach to classifying patients with HF may improve our ability to target interventions, improve patient experiences, and predict outcomes.

Key Question

We sought to develop a new classification tool for predicting patient outcomes and survival time in patients with HF.

Data Source

Dataset: Data from the 2008-2011 Healthcare Utilization Project (HCUP) California Inpatient and Emergency Department (ED) databases were used. These administrative claims databases contain information on 42 million inpatient and emergency visits for over 17 million patients, which are linkable across years.

Inclusion/Exclusion Criteria: We included all patients with at least one admission for HF between 2008-2011. All records for HF patients were retrieved. For training purposes, we used a random sub-set of 822 patients with HF.

Summary Statistics:
Median number of encounters per patient: 3 (IQR 5)
Number of variables evaluated: 168

Approach

Response Variables: Length of stay (continuous), disposition after discharge (categorical), and in-hospital survival time at the last visit on record

Predictor Variables: Patient demographics (age, gender, race, homeslessness, ZIP code level median income quartile), admission characteristics (diagnostic codes, procedures, length of stay, admission type (ED versus inpatient))

Variable Definitions: Predictor variables were grouped into long and short-term features for each patient. Long-term features were summarized across all previous hospital encounters except the last encounter using medians, modes, or counts depending on the variable. Short-term features were similarly characterized using only visits in the 30 days prior to the last visit.

Statistical Approach: Random forests (RF) were used to generate trees that classified patients based on response variables and to identify the relative importance of predictor variables in predicting these responses. Feature selection was intrinsically determined by node impurity using entropy for categorical responses, mean square error for continuous responses, and log-rank for survival time. Model performance was characterized using the out-of-bag (OOB) error rate, which is the proportion of times that a case is misclassified, averaged over all cases. Information from these trees can then be used to develop algorithms for predicting patient response.

Novelty/Innovation

Previous models have used regression techniques with relatively few variables to predict in-hospital survival. We propose a tree-based framework which can be used for both classification and identification of important predictor variables.

Advantages of RF include: 1) They efficiently process large numbers of input variables simultaneously, 2) They do not overfit, and 3) They provide an effective method for imputing missing data and maintain accuracy when a large proportion of the data are missing.

Discussion

Random forests can be used to classify patients with HF based on survival time with a relatively small error rate and to identify the most important variables for prediction. Whereas prior models have yielded C-statistics ranging between 66-75% for in-hospital mortality, our preliminary results here yielded an average misclassification rate of 22%. Next steps include exploration of additional variables and outcomes including length of stay and disposition after discharge. Ultimately, we intend to use these methods to develop prediction models for a variety of outcomes in patients with HF and other conditions.

Limitations: The current approach does not account for the repeated measures design of the data; information is lost by summarizing variables across encounters. We are working on strategies to incorporate the repeated measures design into the random forests for prediction of patient outcomes.

Preliminary Results: Exploratory Analysis

Classification of patients based on in-hospital survival time

(A) Performance of the random survival forests. OOB error rate vs. the number of trees used for computation. After ~400 trees, the OOB error rate stabilized at ~22%.

(B) Variable importance as determined by the random survival forest containing 1000 trees. Blue bars with positive variable importance values indicate the variables contribute to the model. Red bars with negative values indicate the variables do not contribute to the model. Long term admission intensity, patient age, and median length of stay on prior visits were identified as the most important predictors of in-hospital survival time during the last visit.