Decision Tree Based Approaches with Application to a Rabbit Left Ventricular Wedge Preparation Study

Hsu, Yu-Yi, FDA/CDER; Xi, Nan, UCLA; Huang, Dalong, FDA/CDER; Dang, Qianyu, FDA/CDER

Abstract

Background: Drug-induced Torsde de Points (TdP) is a rare but potentially fatal side effect. The risk of TdP is recommended to be assessed at early drug development period. The Comprehensive in Vitro Proarrhythmia Assay (CiPA) steering team published a selection of 28 drugs categorized as low, intermediate and high TdP risk, and set up general principles to quantify models and metrics to be used to predict proarrhythmia risk. Purpose: A type of ex vivo study, rabbit ventricular wedge assay (RVWA), was proposed to be used to assist the early development decision and assist the evaluation of potential risk. Multiple ECG intervals and an ordinal variable representing the early afterdepolarization related incidence are recorded from each sample in a blinded validation study. This research project uses a decision tree based model to predict proarrhythmia based from RVWA results. Methodology: A Bayesian additive regression tree (BART) model was used to analyze data from the RVWA validation study. Each drug is assigned to a risk category based on the predicted probabilities. The model uses a regularization prior to summarize information from multiple binary prediction trees. The posterior probability vectors yielded from MCMC are further summarized and visualized in a distance metric for easier understanding and clearer explanation. Results: The results based on the training and validation framework suggested by CiPA program have 75% (4 out of 16) correct predictions of risk categories. The misclassified drugs are likely to have posterior densities close to boundary of two categories or showing greater uncertainty from the data. The results based on leave one drug out validation has 82% (23 out of 28)correct predictions. Conclusion: The BART model demonstrated robust prediction results. More importantly, the posterior densities provide descriptive information regarding the uncertainty of the categorization. The corresponding distance visualization provides a better understanding of the relationship between a testing compound and reference drugs in different categories.

Introduction

The original classification tree model was proposed by Leo Breiman who used impurity functions to govern the growth of the tree. The method has been extended to multiple variations including ordinal type of response variables, weighted methods, grouped structures and there have been also other different tree structures proposed including the C45 tree that was shown to have good performance especially when there are nominal predictors with more than one categories.

Chipman et al. (2002) proposed Bayesian version of tree models using Bayesian rules for tree structure construction. Chipman et al. (2010) proposed Bayesian additive regression tree (BART) model that further incorporates the randomness in structure using ensemble type of algorithms. The BART model has been applied in multiple areas, including spam email classification, spatial-adjusted missing data imputation (Muller et al., 2007), improving the accuracy of approximating numerical integration (Zhu et al., 2020), etc. There are also many extensions of BART model being developed. The BART model under default prior and settings, as suggested in the original article, is shown to have robust performances. We also use the default settings for our analyses.

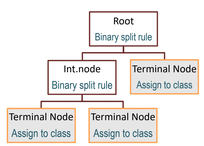
Ex vivo RVWA Study

The study reported by Liu et al. [1] was a blind study tested 34 drugs including 28 with known risk levels using the mechanistically-based rabbit ventricular wedge assay. Each of the drugs was tested for 4 concentrations in 4 wedge preparations. The drug information was blinded to the investigators. The drug information were disclosed after all data analysis and report were completed. The study report was based on the normalized TdP score system at one Cmax. A set of cut-off values were proposed for risk categorization. The TdP score system proposed in the report is easy to use and the relationship between response and explanatory variables can be clearly described.

We considered an alternative approach to analyze the dataset by using the BART model for classification.

Methods

A binary classification decision tree is an approach that recursively split the value space of independent variables and assign each partition a class value. The construction process depends on selected splitting rules, stopping rules and pruning rules. The structure of a grown tree is a directed graph with a root, internal nodes and terminal nodes. At the root as well as each internal node, a rule is selected to improve the prediction based on a pre-specified criteria. A binary tree can be defined by the structure T, as shown below, and a group of parameters corresponding to individual nodes.



Here we use notations similar to those used by Galimberti et al. (2012). Consider observations $\{(y_{di},x_{di1},\ldots,x_{dip});i=1,\ldots,n\}$ where the index $d=1,\ldots,D$, the index $i=1,\ldots,n_d$, the response y_{di} is risk category of the dth drug and ith sample, and $(x_{di1},\ldots,x_{dip})$ is the vector of p predictors correspond to that sample. Assume that the response $y_{di}=w_j\in\{w_1<\ldots<w_J\}$ where $J=3$, and the value $w1$, $w2$ and $w3$ corresponding to low, intermediate and higher risk categories. For simplicity, we use notation $w_j=j$.

Assume that the outcome from the kth tree given tree structure T_k and parameters M_k can be express as $g((x_{di1},\ldots,x_{dip}); T_k, M_k)$ for $k=1,\ldots,m$. The BART model can be expressed as a sum over m trees . The BART model also uses a regularization prior to discount for the complexity of individual trees. To apply for binomial or multinomial outcomes, a data augmentation step was added in the modeling process.

Results and Discussion

Part of the observed data are shown in Figure 1. The three colors in Figure 1 represents group j. The blocks on the diagonal show density plots of individual covariates by group. The off-diagonal blocks are scatter plots of pairs covariates.

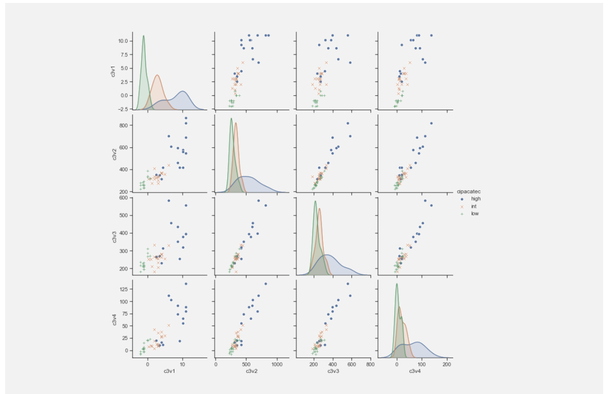


Figure 1. a subset of observed data

As shown in Figure 1, the density plots for individual covariates tend to have certain proportion of overlapping. That makes classification difficult if we use only one or two covariates. The results will be likely to have either high false positive rate or high false negative rate for individual groups.

The modeling results can be summarized in terms of credible intervals. Here we use graphic representation of posterior densities to demonstrate the uncertainty of classification for individual drugs. Figure 2 shows the output of a high-risk drug classification. In this example, the high-risk drug was correctly labeled as high-risk group. The gray dots represent the reference group generated from the training set. The red dots represent a high-risk drug selected from the test set.

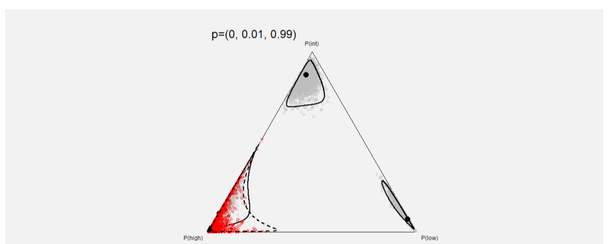


Figure 2. An example of a correctly classified high-risk drug.

Figure 3 shows a misclassified example. The tested low-risk drug was categorized as intermediate risk according to posterior probability. However, the figure shows high uncertainty about the classification.

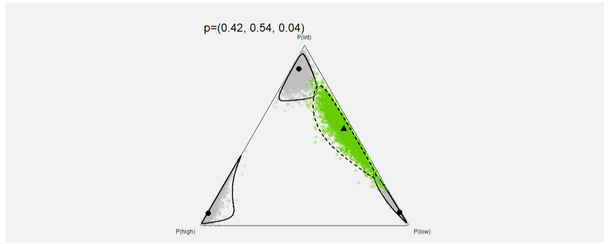


Figure 3. An example of incorrectly classifying a low-risk drug as intermediate risk.

Conclusion

The BART model demonstrated robust prediction results. More importantly, the posterior densities provide descriptive information regarding the uncertainty of the categorization. The corresponding distance visualization provides a better understanding of the relationship between a testing compound and reference drugs in different categories. However, the interpretations of individual covariates from a BART model are not as straight forward interpretations from a usual regression model output while in the out application, being able to understand contribution and impact of individual covariates are rather important for interpretating the results. There were some methods can be used to measure the importance of individual covariates. Further investigations are needed in this area.

Disclaimer

This presentation reflects the views of the authors and should not be construed to represent FDA's views or policies.

Reference

[1] Liu, T., Liu, J., Lu, H.R., Li, H., Gallacher, D.J., Chaudhary, K., Wang, Y. and Yan, G.-X. (2020), Utility of Normalized TdP Score System in Drug Proarrhythmic Potential Assessment: A Blinded in vitro Study of CiPA Drugs. Clin Pharmacol Ther. https://doi.org/10.1002/cpt.2133

[2] Chipman, H.A., George, E.I. & McCulloch, R.E. Bayesian Treed Models. Machine Learning 48, 299–320 (2002). https://doi.org/10.1023/A:1013916107446

[3] Hugh A. Chipman, Edward I. George, Robert E. McCulloch "BART: Bayesian additive regressiontrees," The Annals of Applied Statistics, Ann. Appl. Stat. 4(1), 266-298, (2010)