Random Forests

AUTHOR
Danielle Novinski

PUBLISHED
September 24, 2023

Introduction

This is an introduction to random forests, a machine learning method widely used in a variety of applications. Random forest is an ensemble learning method comprised of multiple decision trees. Decision trees are used for predictions and classification in a variety of artificial intelligence and machine learning tasks. Decision trees divide data by fields, creating subsets called nodes. Cut offs are applied based on statistics. Resampling training data and using multiple trees to reduce bias and variance leads to the random forest approach.(de Ville 2013)

Random forest is a classification and regression ensemble method. Ensemble learning methods use several methods to solve a problem. Diversity among methods is indicated by methods making different errors. This leads to improvements in classifications. Random forest extensions include non-parametric test of significance, weighted voting, dynamic integration, weighted random sampling, online random forest, and genetic algorithms. Applications of random forest include ecology, medicine, astronomy, agriculture, traffic, and bioinformatics.(Fawagreh, Gaber, and Elyan 2014)

The authors suggest modifications to the random forest approach regarding how individual trees are built, the construction of datasets, and how individual predictions are coalesced into one prediction. Conditional inference forests are random forests where the predictors of each split are tested for response and allows adjustment for different predictors. For parameter tuning, if the number of trees increases with the number of predictors, the randomization process will lead to predictors having a good chance to be selected. Predictor selection is important. Some predictors are not informative. The number of predictors is important in order to moderate the effects of strong and weak predictors. The size of trees is determined by the splitting criteria. The size of the leaves is another parameter wherein good predictors could be overlooked if the leaf size threshold is too small. The authors suggest sampling without replacement to avoid bias, and sampling a certain number of observations from multiple classes. Random forest implementations can easily be handled by R packages including randomForest and party. R is a free open source software package with documentation, and includes statistical methods for comparisons. Random forests have been used in several studies related to biology and medicine. In such applications, there is usually a relationship between response and predictor variables and sometimes strong correlations between predictors. Random forests have helped with prediction, ranking of influential variables, and identifying variables with interactions. However, the randomization within implementations prompts the question of reproducability of results. (Boulesteix et al. 2012)

As the size of datasets increases (defined by the number of variables exceeding the number of observations), the performance of statistical methods declines; machine learning methods are preferred. The random forest method is comprised of “weak learners - predictors with low bias and high variance”. Random forest error rates are typically good. The study found that random forest is sensitive to tuning parameters such as number of variables selected and number of trees. Other studies have found good

performance in areas such as predicting customer churn, fraud detection, identifying bacteria and fish, and identifying eye disease. The default value used to split a node is log2 (N) + 1 or sqrt(N) but this may not be optimal. In addition to adjusting the default split value, the authors suggest using a large number of trees, optimizing the number of variables for node splitting per the experiment, using a proximity matrix (checking to see if two values are on the same node) to account for missing values and outliers, and applying feature selection. (Verikas, Gelzinis, and Bacauskiene 2011)

One application of random forests is in the medical field. Lin et al (2019) analyzed a very large dataset in order to determine risk factors for diabetes, a disease affecting 425 million adults where chronic high blood sugar has effects such as damaging organs. This study applied feature selection, where large numbers of attributes are evaluated to identify the most optimal attributes for prediction. The variable importance score was used, as well as ranking the area under curve, in which the highest area under curve is optimal. (Lin, Ji, and Pei 2019)

Another study focused on prediction of students' major and completion status. Supervised statistical learning models consist of response, which is the result to be predicted, and predictors, which are the factors that contribute to prediction. A classification tree can be constructed to predict new results after being trained on predictors. The classifier can divide the data based on conditions and establish groups with many similar observations. A few measures of variable importance can be used to examine the impact of predictor variables. The Gini decrease importance calculates the average impurity measure for predictor across the tree and is easy to use but less reliable. The permutation decrease importance permutes values of a predictor and re-runs the classifier. If the predictor is important, the accuracy will decrease. Thus, the predictors can be ranked to determine the most influential ones. Beaulac and Rosenthal (2019) found that they were able to predict program completion with 78.84% accuracy and predict choice of major with 47.41% accuracy.(Beaulac and Rosenthal 2019)

Methods

The

Analysis and Results

Data and Vizualisation

A study was conducted to determine how...

▶ Code

Statistical Modeling

Conclusion

References

Beaulac, Cédric, and Jeffrey S. Rosenthal. 2019. "Predicting University Students' Academic Success and Major Using Random Forests." *Research in Higher Education* 60 (7): 1048–64. http://www.jstor.org/stable/45217777.

Boulesteix, Anne-Laure, Silke Janitza, Jochen Kruppa, and Inke R. König. 2012. "Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics." *WIREs Data Mining and Knowledge Discovery* 2 (6): 493–507. https://doi.org/https://doi.org/10.1002/widm.1072.

de Ville, Barry. 2013. "Decision Trees." *WIREs Computational Statistics* 5 (6): 448–55. https://doi.org/https://doi.org/10.1002/wics.1278.

Fawagreh, K., M. Gaber, and E. Elyan. 2014. "Random Forests: From Early Developments to Recent Advancements." *Systems Science & Control Engineering* 2 (9): 602–9.

Lin, Shaofu, Wei Ji, and Jiangtao Pei. 2019. "A Method for Selecting Diabetes Features Based on Random Forest." *Journal of Physics: Conference Series* 1237 (2). https://login.ezproxy.lib.uwf.edu/login?url=https://www.proquest.com/scholarly-journals/method-selecting-diabetes-features-based-on/docview/2566217454/se-2.

Verikas, A., A. Gelzinis, and M. Bacauskiene. 2011. "Mining Data with Random Forests: A Survey and Results of New Tests." *Pattern Recognition* 44 (2): 330–49. https://doi.org/https://doi.org/10.1016/j.patcog.2010.08.011.