The Random Forest Algorithm

Benson Ou-yang

Contents

1 Introduction

Machine learning has taken up a storm in the data science community with its complex models and the ability to tackle many different data problems. One of the first definitions of machine learning is programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort (Samuel, 1959). There are two types of machine learning: supervised and unsupervised. Supervised learning is known for classification or regression problems as the goal is to find a relationship in the input data to produce correct outputs. Some examples include using Logistic Regression to filter through spam emails, Linear Regression to predict abalone age or Random Forest to detect ovarian cancer. Unsupervised learning is known as clustering since this type of algorithm tries to group unstructured data.

One of the most popular and powerful machine learning algorithms, Random Forest, is a supervised learning algorithm for classification problems. It is an ensemble machine learning technique. An ensemble is a data mining technique composed of many individual classifiers to classify the data (Abhilasha, 2017). Random Forest consists of many Decision Trees is an ensemble machine learning technique. Decision Trees involve nodes with multiple levels. The input data checks each child node condition starting from the root. The algorithm keeps checking the nodes until it reaches the leaf node, which is the bottom of the tree or until no more agreeable conditions.

2 Preliminaries

The Random Forest algorithm was first developed in 1995 by Tin Kam Ho. Her idea for Random Forest came from the limitations of decision trees where the trees cannot be grown to arbitrary complexity for possible loss of generalization accuracy on unseen data (Ho, 1995). Statisticians from all over the world started expanding on her idea and making extensions to the algorithm. In 1996, a statistician, Leo Breiman, introduced the “bagging” method. This method is performing sampling with replacement, also known as bootstrap aggregating.

Some popular fields like finance, healthcare and e-commerce incorporate Random Forest into providing solutions. Random Forest models can detect fraudulent activities on credit cards, analyze and predict illnesses based on patients’ medical history, or determine how successful a product is.

3 Development

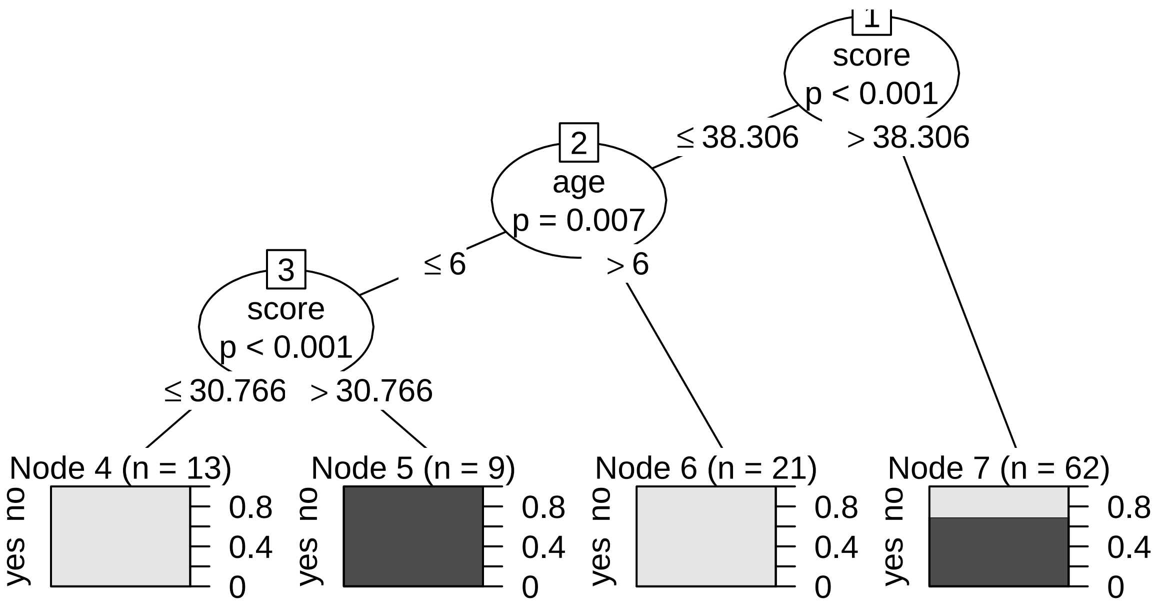

Figure 1: Example of a Decision Tree Produced in R

Figure 1 is from this tutorial for making Decision Trees in R. The tutorial uses a built-in R data set called readingSkills which contains data on subjects reading abilities. The data set contains variables for age, shoe size, reading score, and a binary variable indicating if a subject is a native speaker. The conclusion for this Decision Tree model is that anyone whose reading score is less than 38.3 and age is more than 6 is not a native speaker.

A Random Forest model takes a training set T and forms bootstrap training sets T_k, then constructs classifiers $h(x, T__k) and lets these vote to form the bagged predictor.

Below is Figure 2, which is an image that is taken from Hands-On Machine Learning with Scikit-Learn and TensorFlow by Aurelien Geron. It visually shows the steps of how bagging works.

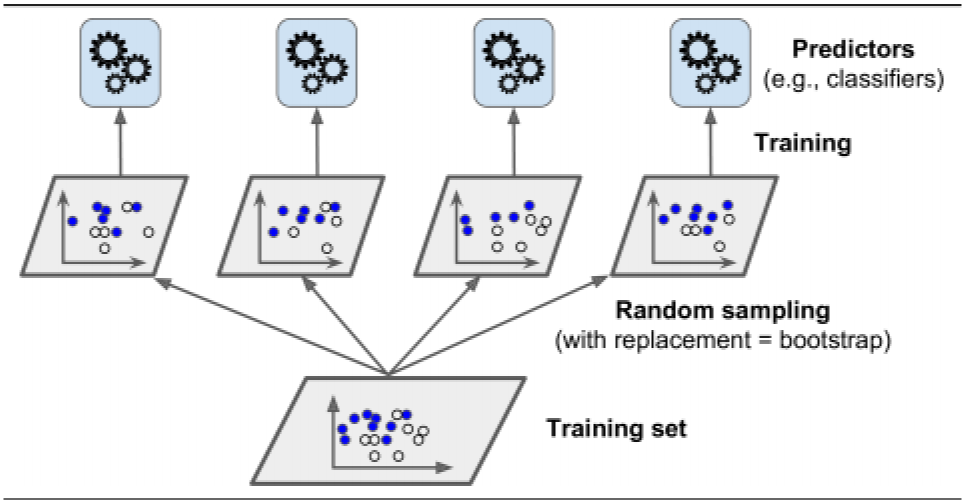

Figure 2: Visual interpretation of bagging

A Random Forest model randomly selects a small group of input variables to split on at each node. Using CART methodology can grow the tree to maximum size (Breiman, 2001). CART is known as Classification and Regression Trees. CART builds trees by recursively splitting the variable space based on the impurity of the variables to determine the split till the termination condition is met (Santhanam and Sundaram, 2010). The Random Forest algorithm takes random subsets of features and uses CART to grow a great diversity of Decision Trees. A greater tree diversity trades a higher bias for a lower variance, resulting in an overall better model (Geron, 2019).

4 Testing

A good machine learning model requires high accuracy. A way to achieve that is to select the right features. There are methods such as forward, backwards or stepwise selection based on the Akaike information criterion. Using these one variable elimination or addition methods is a "greedy search technique" (Visa, Ramsay, Ralescu, Knapp, 2011). Using a confusion matrix is a better method for feature selection. A confusion matrix is a size n x n matrix that outputs the prediction accuracy and classification error based on the correct and incorrect positive and negative predictions. Confusion matrix-based attribute selection constructs subsets of attributes that have good individual classification power and are complementary. Another tactic of selecting variables is using the variable importance measurement. The variable importance ranks each variable in the model on how well they contribute to the model.

Out-of-bag classifiers are any votes for the classifiers for bootstrap training sets T_k that do not contain y,x we aggregate them. The out-of-bag estimate for the generalization error is the error rate of the out-of-bag classifier on the training set (Breiman, 2001). The generalization error is a measure of how accurate an algorithm can predict values given new data. Statisticians in the past have proposed different methods of using out-of-bag estimates to estimate the generalization error. Breiman's paper in 1996 provided evidence that shows that using out-of-bag error estimates is as accurate as using a test set of the same size as the training set.

5 Discussion

Random Forest has many advantages over other classification algorithms. The ability to grow trees due to randomness does not allow the trees in the model to overfit. It gives out estimates like the out-of-bag estimate to measure error, strength, and correlation and outputs the importance of variables (Breiman, 2001). Statisticians have been researching extensions and additions to the Random Forest algorithm to improve accuracy and performance. ReliefF Random Forest is a method that evaluates variables to select subsets of variables based on a quality metric which helps decrease correlation while maintaining strength. Another method is Dynamic Random Forest which only reliable trees are allowed to grow. Since Random Forest adds trees independently, that can affect the performance of the forest. Dynamic Random Forest prevents that by using a subset of trees as the base classifier.

Many fields incorporate the Random Forest algorithm into solving problems. This research paper by A. Arfiani and Z. Rustam took the bagging and Random Forest method for classifying ovarian cancer data. The result for their Random Forest model reached 98.2% accuracy, while the bagging method reached 100% accuracy on 90% of the training data. Another instance of the Random Forest algorithm used is in a financial fraud detection model. The researchers experimented with parametric and non-parametric models and found that the Random Forest algorithm produced the highest accuracy. Random Forest can also predict the direction of stock market prices. The authors L. Khaidem, S. Saha, and S. Dey found that the Random Forest algorithm performed better than other algorithms in predicting stock prices. They also thought that the out-of-bag estimates were helpful.

6 Conclusion

The Random Forest algorithm is a popular methodology due to its many features and high accuracy output. The out-of-bag estimate for the generalization error measures how accurate the algorithm can predict given new values. Variable importance is a feature that is also useful in the variable selection of the model. Using a confusion matrix can also determine prediction accuracy and classification error based on correct and incorrect positive and negative predictions. Over the years, statisticians and researchers have been studying the Random Forest algorithm and making extensions. Some extensions include ReliefF Random Forest that selects subsets of variables based on a metric. Dynamic Random Forest is another extension that uses a subset of trees as the base classifier instead of one tree. Many different fields used Random Forest as it outperformed other methodologies. Some cases include detecting ovarian cancer, financial fraud detection and predicting stock prices. Overall, the Random Forest algorithm produces highly accurate models and has measurements to minimize error and maximize accuracy.

7 References

[1.] L. Breiman, Random Forest, 2001

[2.] E. Goel and Er. Abhilasha, Random Forest: A Review, 2017

[3.] A. L. Samuel, “Some studies in machine learning using the game of checkers,” in IBM Journal of Research and Development, vol. 44, no. 1.2, pp. 206-226, Jan. 2000, doi: 10.1147/rd.441.0206.

[4.] A. Arfiani and Z. Rustam, “Ovarian Cancer Data Classification Using Bagging and Random Forest”, AIP Conference Proceedings 2168, 020046, Nov. 2019

[5.] D. Bhalla, Complete Guide to Random Forest in R, 2015

[6.] N. Donges, A complete guide to random forest algorithm, 2019

[7.] C. Liu, Y. Chan, S. Kazmi, H. Fu, “Financial Fraud Detection Model: Based on Random Forest”, 2015

[8.] T.K. Ho, “Random decision forests,” Proceedings of 3rd International Conference on Document Analysis and Recognition, 1995

[9.] T. Santhanam and S. Sundaram, “Application of CART Algorithm in Blood Donors Classification”, 2010

[10.] L. Khaidem, S. Saha, S. Dey, “Predicting the direction of stock market prices using random forest”, 2016

[11.] Decision tree in R Tutorial

[12.] Random Forest in R Tutorial