CS6140 - Final Project Proposal
Brian Sherrill, Chuong Tran, and Mark Grube
03/31/2017

Investigating Three Different Random Forests Related Methods

Methods: CART (specifically Regression Tree), Rotation Forest, Online Random Forest

We are looking to implement different methods of Random Forests (RF) found in three different research/scientific papers and apply them to the same datasets to evaluate how good or accurate they are. The three methods are Classification and Regression Tree (CART) but more specifically toward Regression Tree (RT) part, Rotation Forests, and Online Random Forest.

The first one we are looking at is the Regression Tree (RT) part of CART since the response variables from the dataset are mostly numeric or continuous. Plus the goal of RT is to predict a certain outcome which serves as a good starting point for our project. Besides, RT is also one of the advanced topics that has not been introduced properly in class. The idea behind RT is to fit a regression model to the target variable which usually does not contain classes by taking into account of each of the independent variables. Then we will compute the Sum of Squared Errors (SSE) for each split point, compare them and the one that has the lowest SSE will be chosen as the root node or split point later. The process of RT is very interesting in the fact that it is applicable to prediction instead of classification. Moreover we get used to dealing with the response variable only has two categories or multiple one using standard Classification Tree or C4.5 implementation. As a result, when it comes to the response variable is continuous, they will fail and RT is a better choice in this case. More importantly, it can fit and be used in a lot of traditional statistical models as well. Most of the topics and methods we cover in class do not mention a lot on how we can deal with this type of variable. That is also the main reason why we choose RT and want to introduce it to the class as it provides a good learning point as well.

Based on the International Statistical Review journal, Professor Wei-Yin Loh has offered his help in reviewing a lot of Machine Learning and Statistical techniques and proposed methods that other scientists have done over the years for CART. Therefore it will serve as a good platform for us to start looking into it and see what others have done and see if we can replicate that. One of the methods that we are particular interested in is to use RT introduced by Professor Breiman and others in 1984 as it has been well-received and also one of the first as well. Plus he also mentions on how RT can be coupled with RF to use for regression analysis (or nonlinear multiple regression). RF can help to improve the accuracy and instability of the tree.

The second Random Forest method that we will look to implement is the Rotation Forest. The rotation forest is an ensemble of decision trees. The algorithm works by splitting the

dataset into K random subsets. Then we perform K axis rotations on the features to improve diversity and accuracy within the ensemble, since decision trees are sensitive to rotations on the feature axis. We use bagging to combine the results of the decision trees. The authors of the paper claim that Rotation Forests perform better than other types of random forests, such as AdaBoost. We would like to see if this claim holds up, and compare the performance of Rotation Forests to the other types of Random Forests we have chosen to study.

The third Random Forest method we would like to examine is on-line random forests. This algorithm would be relevant to situations where you data is coming in over time, or that your data is changing. The algorithm uses at temporally weighted scheme, so that newer data is given a higher weight. This is interesting because we live in a constantly changing world, where we would like to be able to predict based on what recently happened, as opposed to stale data. Furthermore the paper states that the algorithm converges to off-line RF, which would be something we could test empirically with our own data set.

This will be an interesting project as we have not done these methods in class. Initially we should examine each algorithm individually to gain better insights into each algorithm. Afterwards we can compare and contrast advantages and disadvantages, as well as discuss in what situation each algorithm would make sense to use. Implementing these algorithms will give is a deeper understanding of not just RF, but different techniques in machine learning. The algorithms, and the experience we gain, would be useful not only in future interview, but in our future careers.

Datasets:

Since we are looking at the performance of multiple implementations of Random Forests, we are going to use a couple datasets and run them on each to see how they perform. Some of the datasets we will use are:

Fantasy Football Rankings -

The dataset we all agree on is from the pro-football-reference.com and the section we interest in is the fantasy rankings from the year of 2007 as it provides the more complete version of the dataset, comes with csv file, easy to modify and contains numerical continuous data. It also can make an engaging presentation as well. The most important things that we look in a dataset are it is big enough to perform on and free to use.

Boston Housing -

Besides, we also have a backup as well using the Boston housing dataset from the UCI Machine Learning Archive, but compare to the first, it is not very interesting to look at and some of the data needs to be refined. Therefore, we agree to give the first dataset a try since we want to focus more on studying the methods first before worrying about the dataset.

Reference:

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.150.1671&rep=rep1&type=pdf
http://ieeexplore.ieee.org/abstract/document/1677518/
http://www.stat.wisc.edu/~loh/treeprogs/guide/LohISI14.pdf
Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). Classification and Regression Trees. Belmont: Wadsworth.