Final

Conner McCloney

# Introduction

In 2011, the High Time Resolution Survey was conducted. This survey was a digital, all-sky survey designed specifically to scan the night sky for potential pulsars (or pulsar candidates), and then determine their validity. A pulsar is a rare type of rotating Neutron star that emits electromagnetic radiation, and can only be measured when this beam is aimed directly at Earth. These pulsars were searched for using large radio telescopes looking for periodic radio signals similar to what a pulsar would produce. Many measurements taken that could be a pulsar, known as a candidate, were recorded, however in practice most observations recorded are caused by radio frequency interference (RFI) and noise. This data set records the mean, standard deviation, skewness, and excess kurtosis of the integrated profile, which is an array of variables that describe the signal recorded, and of the Dispersion Measure - Signal-to-Noise Ratio (DM-SNR) curve. The DM-SNR curve describes the relationship between the two for the observed signal. A curve whose SNR peaks at a DM of zero is likely RFI, whereas a legitimate signal would be expected to peak at a DM greater than zero. Lastly, there is the target\_class variable, which is a binary classification variable of whether the given signal was truly a pulsar or not. I am interested in investigating which, if any, of the characteristics of the DM-SNR curve and/or the integrated profile are useful in accurately classifying a pulsar candidate, and how accurate a model of this sort is.

# Methods of Analysis

From this data set, data for eight variables in total were collected for every observation: the mean, standard deviation, skewness, and excess kurtosis of both the DM-SNR curve and the integrated profile. I chose to conduct two different methods of classification on this data, support vector machines (SVMs) and tree-based methods. SVMs are a binary classification method, and they work by taking the data and plotting it in -dimensional space, where is the number of predictors present, and attempt to separate the data into two classes using a ‘hyperplane’, or a flat affine subspace of dimension -1. This hyperplane is designed to maximally separate the data, meaning that the orientation and location of the hyperplane is chosen such that the width of the margin between the hyperplane and the points on either side of it is maximized. This method can be altered to allow points of a class to fall on the wrong side of a hyperplane if the data is not separable, or to prevent overfitting of the training data. Also, SVMs can be extended by adding a kernel, which is the inner product between two training data points, to allow the hyperplane to be non-linear.

Three tree-based methods were used on this data: bagging, Random Forest (RF) and boosting. Tree-based methods create a series of decision rules that result in two splits based on the characteristics of the data, until a classification is made on a data point when it reaches a leaf-node of the tree. Bagging is a method also known as bootstrap aggregation that functions by aggregating the results of separate training sets, and average them in order to obtain a single model with low variance. The accuracy of this method is measured using Out-of-Bag error estimation, which measures the test error of the model generated. Random Forests are an improvement over bagging theoretically, because they tweak the method in an attempt to decorrelate the trees created. The same method as bagging is used, but when building a decision tree, only a random sample of predictors is chosen to create the model of all predictors. Doing so ensures the large reduction in variance that bagging claimed to provide, but may not in certain situations, such as when there is only one strong predictor in a data set, with a few moderately strong ones. Lastly, there is the method known as boosting, which also works similar to bagging, except each tree is grown sequentially, using information from previously grown trees, and instead of bootstrap sampling, each tree is fit on a modified version of the original data set.

# Summary of Statistical Findings

|  |  |
| --- | --- |
| Method | Misclassification Rate |
| SVM – Linear Kernel | 0.021 |
| SVM – Radial Kernel | 0.039 |
| Bagging | 0.028 |
| Random Forest | 0.029 |
| Boosting | 0.019 |

First, beginning with SVMs, I analyzed the data using two kernels. The first kernel was linear, which simply makes the hyperplane a linear boundary between the two classes. Secondly, I used a radial kernel which allows the boundary to be completely non-linear. Next, I ran each of three tree-based methods on the dataset. The confusion matrices and out-of-bag error estimate for Boosting can be found in the Appendix.

*Table 1*

From these results we can see that among the two SVMs, the linear kernel-based approach actually performed slightly better, although both methods had a misclassification rate less than 5% (Linear: 2.1%, Radial: 3.9%). Among the tree-based methods, Boosting performed the best, with a misclassification rate of only 1.9%, relative to Bagging and Random Forest which were 2.8% and 2.9% respectively. All five methods of analysis performed roughly equivalent, although the best of all methods was the tree-based Bagging approach. The rates are shown above (Table 1).

# Conclusion

It is interesting to note that the Boosting approach also measures the relative influence of each predictor present in our model, which is a measure based on the number of times that predictor is selected for splitting weighted by how much it improves the model. From the output below (Figure 1), we can see that the Excess Kurtosis of the Integrated Profile is given a very large relative influence value, whereas all the other predictors are given relatively small values. This may be due to the fact that since it is a measure of the integrated profile, it shares information with other predictors of the integrated profile in the model, so that only one of these predictors is truly necessary to make a good split. However, this does not account for why predictors based on the DM-SNR Curve have such low relative influence values as well.

Based on the results above, we can see that every method attempted performs well on the data. This was done after subsetting the data, and testing the accuracy of it on the majority of our data, known as our test data set. Although the misclassification rate is quite small at ~2-3%, that still roughly equates to ~300-400 misclassified observations in our dataset of = 16,648, which is quite a lot. This says that there is still room for improvement to be made, even in our best approach of Bagging. Possible improvements could also be made in the data collection phase as well, incorporating more measures of the Integrated Profile and DM-SNR Curve, or new measurements of these pulsar candidates entirely if possible.

# Figures

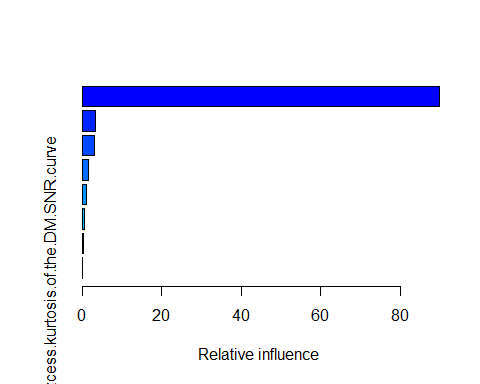


Figure 1