Bagging and Random Forest Classification: Wine Dataset

Raj Choudhary*
**School of Computer Science and Engineering, VIT Chennai, Tamil Nadu, India 600127*
*Email: raj.choudhary2016@vitstudent.ac.in**

***Abstract*: Bagging and Random Forest Classifiers are used to classify the wine based on the alcohol content and OD280/OD315 of diluted wines values. We also the compare the result of each of the above-mentioned algorithms with simple Decision Tree Classifier that is also used to classify the wine samples based on the mentioned attributes. Here we observe that even though all the classifiers have good train accuracies, the simple Decision Tree Classifier is not showing good results in the test set as it is not able to generalize the pattern in the training set.**

***Index Terms* – Decision Tree Classifier, Bagging Classifier, Random Forest Classifier, wine dataset.**

1. Introduction

In this paper, wine classification is attempted using wine dataset which contains the different data about the chemical analysis of wine grown in the same region in Italy but derived from two different cultivators. For the analysis, data from two cultivators is used so that it is better able to visualize the result. Also, the classification is result is compared between simple Decision Tree Classifier, Bagging Classifier and Random Forest Classifier.

2. Methodology

The wine dataset is collected from the https://archive.ics.uci.edu/ml/machine-learning-databases/wine. It contains data that are results of chemical analysis of wine from the same region of Italy. The dataset is split into training set and test set. Then it is passed to the different classifiers that classify the result based on the attributes Alcohol content and OD280/OD315 of diluted wines values.

3. Dataset – Wine dataset

The data in the dataset are the data from the chemical analysis of the wine samples collected from the same region in Italy but derived from three different cultivators. The analysis determined the quantities of 13 constituents found in each of the three types of wines.
The attributes are:

- Alcohol
- Malic acid
- Ash
- Alcalinity of ash
- Magnesium
- Total phenols
- Flavanoids
- Nonflavanoid phenols
- Proanthocyanins
- Color intensity
- Hue
- OD280/OD315 of diluted wines
- Proline

4. Analysis and Result

From performing the classification on the dataset using the above algorithms and visualizing the result, we can see that all the algorithms perform best on the training set with an accuracy score of 100%. While Decision Tree Classifier is not able to generalize the pattern to the test set, it obtains a test accuracy score of 83.3%. But the two other algorithms perform better on the test set as compared to the Decision Tree Classifier. Both Bagging Classifier and Random Forest Classifier perform better on the test set with an accuracy score of 91.7%. This shows that as ensemble algorithms used a number of classifiers of the same type to generalize their result, the perform better than the base classifier used which in this case is Decision Tree Classifier.

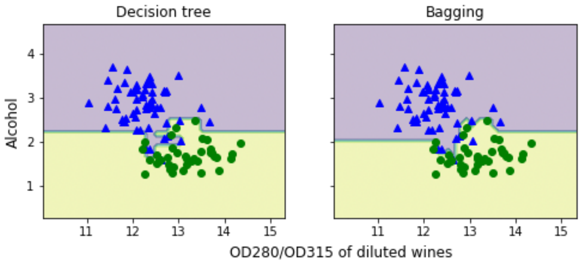

Figure 1. Classification by Decision Tree and Bagging Classifiers

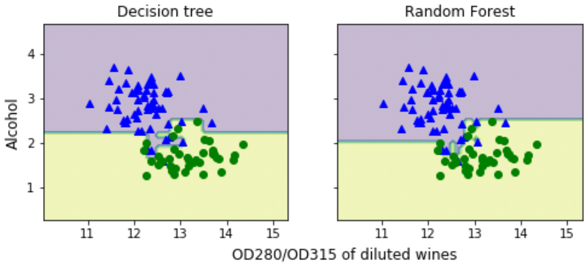

Figure 2. Classification by Decision Tree and Random Forest Classifiers

5. Conclusion

Hence we discovered that ensemble learning algorithms are better in performance than their base classifiers. They are able to perform better than the base classifier because they generalize their results by using a number of base classifiers that acts as their hyper parameter. Moreover, the performance of these ensemble learning algorithms can further be increased by performing tuning of the hyperparameters using GridSearchCV.

6. Reference

[1] Wine recognition data: Updated Sept 21, 1998 by C.Blake : Added attribute information https://archive.ics.uci.edu/ml/machine-learning-databases/wine