# Ensemble methods

## Introduction

As the ensemble method used in building classifiers successfully. Two problems about ensemble method appeared. One is how to create diverse classifiers. The other one is how to establish the weights that individual classifiers contribute to the result. In this assignment, I will introduce two ways of solving those problems and discuss the advantages and disadvantages of those two approaches firstly. Then, a plan made by myself be introduced to solve problem of the weights of contribute to ensemble’s result. In addition, reason of why developed method better than the original one by preformed in the end.

### First approach

There are two existing ways of ensemble method, which are adaboost algorithm and random forest. Firstly, adaboost algorithm is a kind of boosting method. For boosting method, it like to train many weaker learner, which can only do binary classification. Then, combine those weaker learner to strong learner. Adaboost algorithm is very similar. At the beginning, the same weight distribution is given to each sample. Then adaboost algorithm start to train a classifier then evaluate the error of classifier. After that, algorithm determine the weight of classifier based on error of classifier and update the distribution of sample. In the end. Combining every weaker learner which be trained by algorithm. (Ensemble Method Foundations and Algorithms, 2012 p.23) In particular, if classifier get lower error at training process, the higher weight it will get. The advantages of this algorithm are high accuracy and prefect weight distribution of each classifier, because the weight is be distributed based on the accuracy of each classifier when training. In addition, it can use different classifier as weaker learner. However, there are 2 disadvantages of this algorithm. One is long running time. Because, it has two weight distribution need to change in training process which are weight of classifier and weight of sample. Another disadvantage is the difficult of choosing the number of weaker learner.

### Second apporach

Secondly, random forest is a kind of bagging algorithm. The bagging algorithm focus on training many independent base learners.(Ensemble Method Foundations and Algorithms, 2012 p.48) Then combine them together. Random forest is one sample of bagging algorithm. At beginning of random forest, a dataset is divided into two parts which are training dataset and test dataset. Between those two datasets, training dataset is smaller than test dataset. Then training dataset are used to build many random decision trees. After that, random forest algorithm combine those random decision trees together to get result. There is one thing need to be noticed, the weight of each random decision tree generated by random forest is same.( Ensemble Method Foundations and Algorithms, 2012 p.48) The advantage of random forest are high accuracy and low generalization error. According to Zhi-Hua(2012, p.47), error can be significantly reduced by combining independent base learning. Random forest is a kind of parallel ensemble method. Zhi-Hua(2012 p.47) also said that the basic motivation of parallel ensemble methods is to exploit the independence between base learners. So, that is the reason of why random forest can get low error.

However, there is one disadvantage of random forest which is the overfitting problem when dealing some dataset with large anomalies.

## Personal plan

In this part, I will brief introduce my plan about solving the choosing of classifier and weight of each classifier. For the choosing of classifier, I deicide to use bagging methods to get many independent classifiers. In addition, those independent classifiers are random decision tree. For the weight of each classifier, I decide to use boosting method to give different weight based on the accuracy in training process. If the classifier get lower error in training, it can get higher weight. In the end, the result will be given based on the voting of different classifiers with different weights

## Reasons of why new plna is better

n this part, I will point the reason why new plan better than conventional one. There are three reason. Firstly, new plan can deal all kinds of dataset. Because, new plan use decision tree as individual classifier. Decision tree can deal both continuous data and discrete data. Secondly, new plan has lower generalization error, because, every classifier is independent. The combination of those independent classifier can reduce generalization error. Thirdly, it increase the accuracy through set different weight based on error at training process.

## Conclusion

In conclusion, this assignment shows two existing ensemble methods which are random forest and adaboost. Then, the advantages and disadvantages of those two methods are indicated. The advantages of adaboost are high accuracy and prefect weight distribution. The disadvantages are long running time and hard to choose number of classifier. The advantages of random forest is low generalization error. The disadvantage is overfitting problem. After that, a personal plan is shown to address issue of the choose of individual classifier and weight problem. In the end, 3 reasons are given to show why new plan is better than conventional one.

## Reference

Ensemble Method Foundations and Algorithms 2012, Zhi-Hua Z., Taylor & Francis Group, LLC