STAT745 Term Project - Part 3

Multi-class Dataset: GLASS

Dec 10, 2010

Abstract This part of the project deals with a multi-class dataset called GLASS dataset. We will use random forests to predict the classes and we will observe the iteration plots. We use iteration plots to find the optimum parameters, i.e. the optimum tree size (how many trees) and the number of randomly selected features. We found that the optimum number of trees to be used is 200 and the optimum number of random variables to be used for splitting is 1. After finding these values, we evaluated the variable importance according to mean decrease in accuracy and mean decrease in Gini index. We evaluated the importance of variables over all variables as well as for each variable separately. As it turns out the variable importance per every individual class differs. However, overall variable importance w.r.t. accuracy and Gini index have common properties to be reported: The top 3 variables w.r.t. accuracy and Gini are common. The common top 3 most important variables are: Refractive Index, Magnesium (Mg) and Aluminum (Al).

1 Introduction

In the third part of the project we are going to be dealing with a multi-class dataset: *"glass.dat"*, a.k.a. glass dataset. Before going into the analysis part with random forests, I manually inspected the data to see if there are any missing values or anamalies. Fortunately there were no missing values inthe dataset. Glass dataset consists of 11 variables, among which 10 variables are independent variables that describe characteristics of the glass and the 11^th variable is the dependent variable, which tells the type of the glass, i.e. the class value. Furthermore, glass dataset consists of 214 instances. Note that the first variable is the ID of the instances, which does not tell any information. However, it may cause some confusion during the learning, because the learner may treat it as a variable that actually tells something about the instances. Therefore, I did not include the ID variable during predictions.

Ekrem Kocaguneli
Lane Department of Computer Science and Electrical Engineering
West Virginia University
Morgantown, WV 26505, USA
E-mail: ekocagun@mix.wvu.edu

The experimental part of consisted of 2 parts. In the first part we try to find the optimum parameters for random forests. In the implementation of random forests in R, there are two parameters that will play a critical role on the performance: The number of trees and the number of variables randomly selected for the splitting at every node. We first found the optimum number of trees to be used as 200 and they we used this value to find the optimum number of variables to be used. The optimum number of variables to be used was 1. In the second part of the experiments, we wanted to see the variable importance. The R random forest implementation provides a function called importance, which returns the variable importance in terms of mean decrease in the accuracy and the mean decrease in the Gini index. It returns those values both for the overall classification and for every class separately. When we order these variables w.r.t. mean decrease amounts, we saw that each class behaves differently and favors different variables, whereas for the overall classification the top 3 variables selected by accuracy and Gini were the same. Below are the details of these results.

2 Results

In this part, we are going to use random forests and apply it on the dataset to see the change in the iteration plots, i.e. we will observe the performance of random forests on glass data with changing number of trees and changing number of randomly selected features. So as to find the optimum parameters, I first started with the number of trees and plotted the error percentage (percentage of wrongly classified instances) with respect to the number of trees. Below in Figure 1 we see the error percentages for different number of trees from $n = 1$ to $n = 500$. Note that in that scenario, we use the default value for the number of randomly selected variables at each split, which is the square-root of the number of variables. See in Figure 1 that after about the tree number of 200, the error rate stagnates and the increasing number of trees dos not improve to reduce the error rate. Therefore, I will use 200 as the optimum tree amount for the GLASS dataset.

Another parameter we need to decide is the number of parameters that are randomly chosen for random forests. The randomForest package in R allows us to specify that variable by using *"mtry"*. Notice in Figure 2 that different mtry values usually give quite similar results, i.e. almost all the mtry values stagnate around the error rate of 0.2. Only for $mtry = 1$, the error rate slightly goes under 0.2, which is better than the other values. Therefore, we will use $mtry = 1$ as the optimal value.

Another property we will be interested in will be the importance of variables. I used the $randomForest$ package of R, for that question and this package allows the user to see the importance associated with each variable via *"importance"* routine. The importance routine lets us know the importance of the variables according to 2 different criteria: Gini Index and mean decrease in accuracy. Gini index is a simple function used to measure the statistical dispersion of a distribution. It is usually used to assess the inequality in income or wealth. Its value can change between 1 (one instance/person takes all) and 0 (every instance/person receives the same amount). For both decrease in accuracy and decrease in Gini index, we want number as big as possible, i.e. the greater the decrease in any of the two variables, the more important the variable is. Note that in Figure 3 there are the total decrease amount coming from each variable (the last two columns) and also there are per-class based accuracy decrease amounts due to different variables.

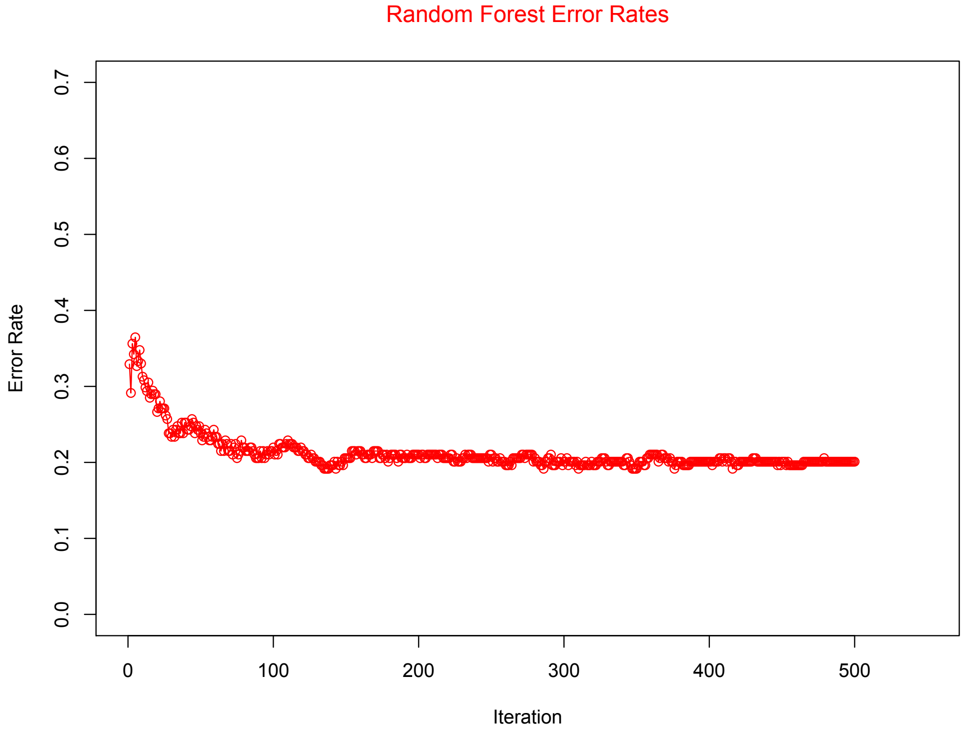
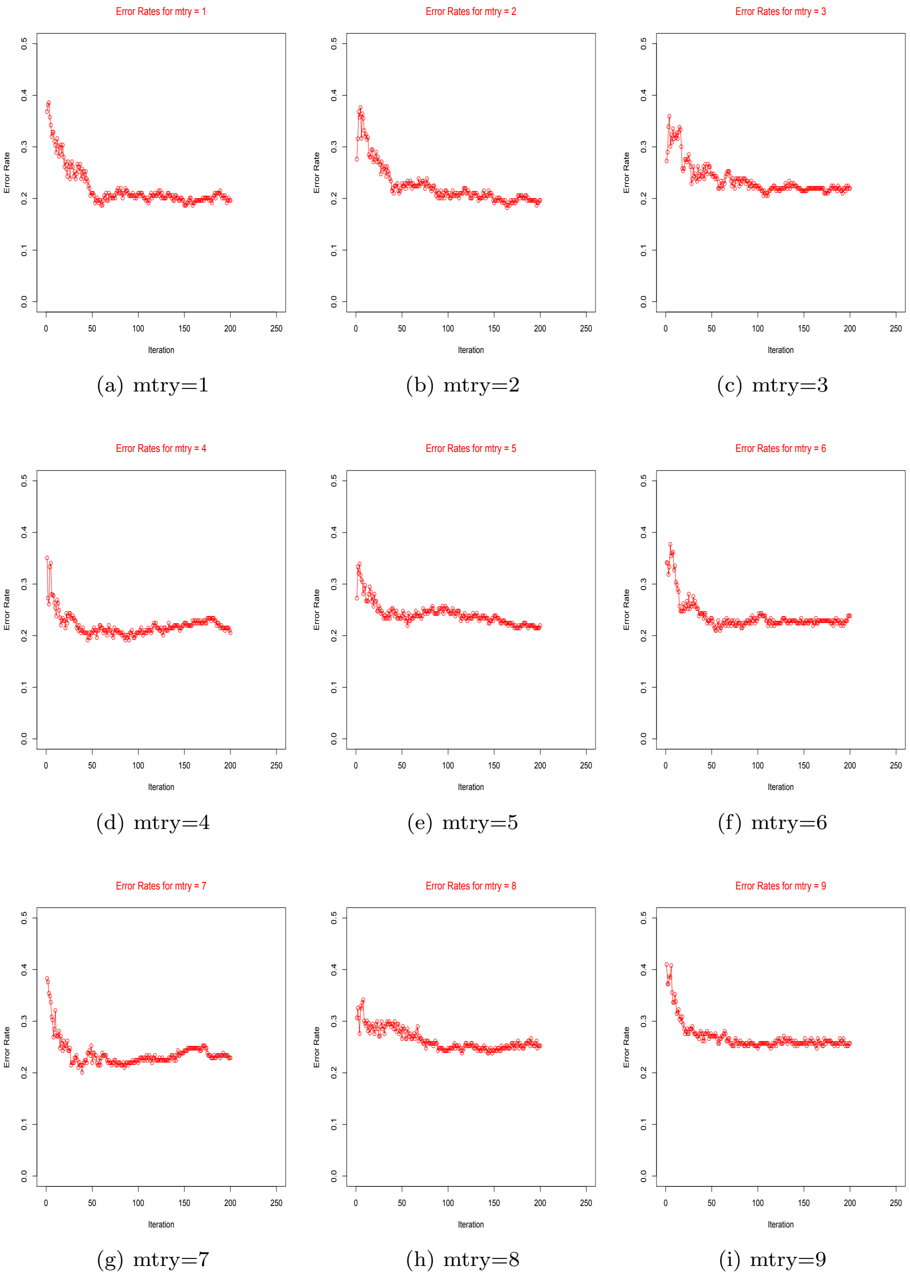


Fig. 1: The error percentages for random forests with different number of trees, from 1 to 500. The error rate is pretty much stable after the tree size of 200, therefore we will use 200 as the optimum tree size in random forests from now on.

When we sort the columns of Figure 3 from highest to lowest separately, then we get the order of importance for the variables for each variable and evaluation criterion. In Figure 4 the result of the ordering of variables with respect to each variable or with respect to accuracy or Gini index is provided. Note that for each variable there are different rankings of variables and it is difficult to claim a most effective variable. However, when we look at the top 3 attributes according to accuracy and Gini index, we see that the top 3 variables are common. They are: RefInd (Refractive Index), Mg (Magnesium) and Al (Aluminum).



(a) mtry=1

(b) mtry=2

(c) mtry=3

(d) mtry=4

(e) mtry=5

(f) mtry=6

(g) mtry=7

(h) mtry=8

(i) mtry=9

Fig. 2: Different mtry values and the error rates for random forests consisting at most 200 trees. The x-axis in every graph is the number of trees and the y-axis is the error rate. Note that all the random forests are usually above 0.2 error rate, whereas only for the case of $mtry = 2$, the error rate can go under 0.2. Therefore, we will pick up $mtry = 1$ as the other optimal parameter.

Variable	Class1	Class2	Class3	Class5	Class6	Class7	MeanDecreaseAccuracy	MeanDecreaseGini
X.RefractiveIndex	3.67	2.89	4.45	2.77	3.84	1.78	2.14	22.03
X.Na	1.83	1.78	0.75	4.85	6.61	3.54	1.73	16.22
X.Mg	3.5	3.27	4.18	8.43	8.19	4.25	2.19	24.96
X.Al	3.52	2.78	3.08	6.43	-0.76	4.12	2.11	24.96
X.Si	2.56	1.93	1.45	1.96	2.78	1.18	1.74	13.64
X.K	2.61	2.21	1.7	3.53	9.09	2.21	1.85	14.27
X.Ca	2.6	3.27	1.76	6.28	0.16	1.64	2.06	20.58
X.Ba	1.2	2.07	1.93	2.79	4.95	5.61	1.93	13.6
X.Fe	0.65	0.62	0.37	0.19	2.76	1.38	0.67	6.74

Fig. 3: The importance of variables with respect to mean decreases in accuracy and Gini index.

Class1	Class2	Class3	Class5	Class6	Class7	MeanDecreaseAccuracy	MeanDecreaseGini
X.RefInd	X.Mg	X.RefInd	X.Mg	X.Al	X.Ba	X.Mg	X.Mg
X.Al	X.Ca	X.Mg	X.Al	X.Ca	X.Mg	X.RefInd	X.Al
X.Mg	X.RefInd	X.Al	X.Ca	X.Fe	X.Al	X.Al	X.RefInd
X.K	X.Al	X.Ba	X.Na	X.Si	X.Na	X.Ca	X.Ca
X.Ca	X.K	X.Ca	X.K	X.RefInd	X.K	X.Ba	X.Na
X.Si	X.Ba	X.K	X.Ba	X.Ba	X.RefInd	X.K	X.K
X.Na	X.Si	X.Si	X.RefInd	X.Na	X.Ca	X.Si	X.Si
X.Ba	X.Na	X.Na	X.Si	X.Mg	X.Fe	X.Na	X.Ba
X.Fe	X.Fe	X.Fe	X.Fe	X.K	X.Si	X.Fe	X.Fe

Fig. 4: The importance of variables with respect to mean decreases in accuracy and Gini index.