471-Final analysis report

By Guangjin Zhou

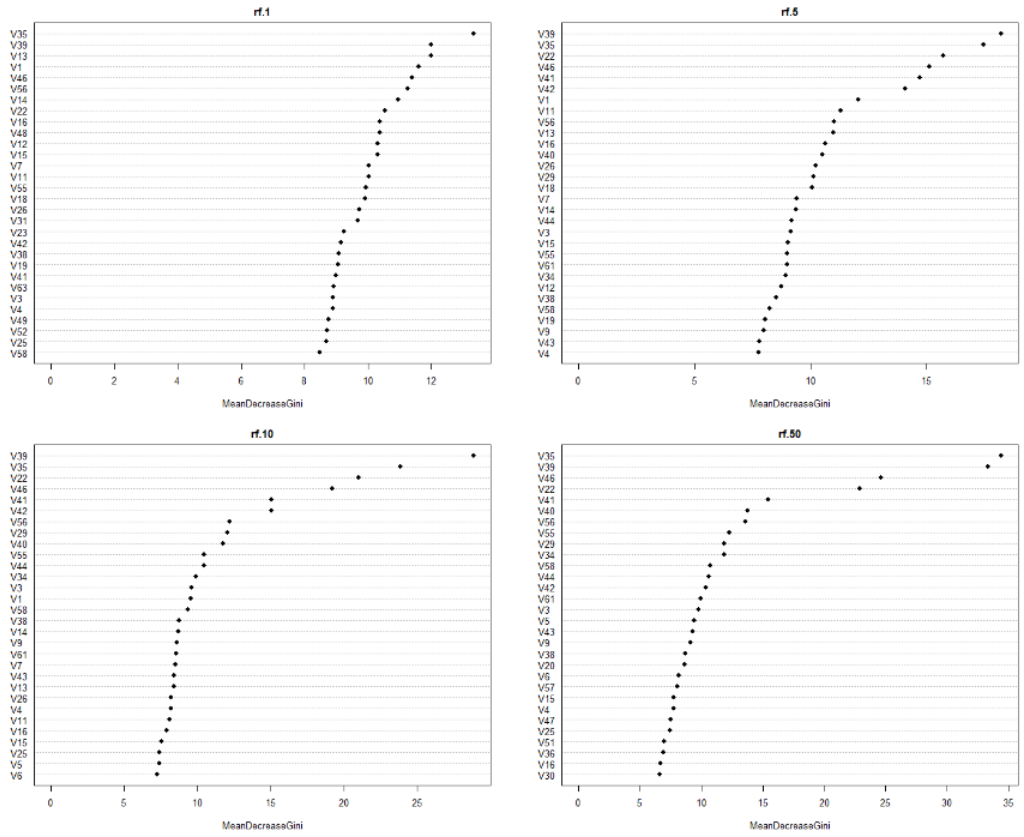
I explored the bankruptcy dataset by three different supervised learning approaches and two unsupervised learning approaches. To run the program smoothly, I run the five methods separately. My 5 RMD codes files and 5 html outputs were uploaded at GitHub account, and following description is my brief summary report.

1 Random Forest:

1) By converting binary outcome as factor, I fitted four RF models with 200 trees and M hyper-parameter as 1,5,10 and 50.

2) I learned two different plots to demonstrate the variable importance. By comparison of predicted error rates, I concluded that M hyper-parameter as 50 gave best prediction model, but the three models with M as 5,10 and 50 do not make big difference (see Figure below).

3) The best tree can be obtained after fitting about 50 trees with these models.

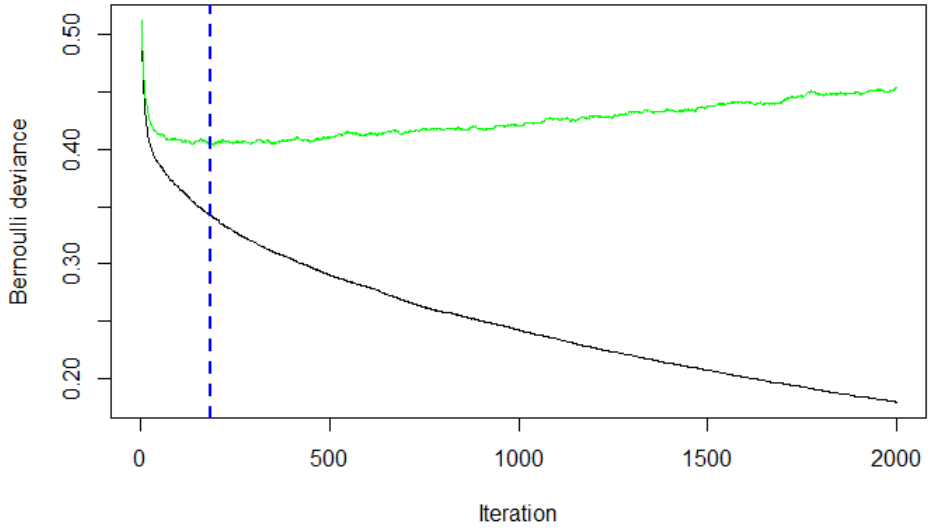


2 Gradient boosting:

1) Two gradient boosting models with 200 and 2000 trees fitted respectively, both models have same relative influence number-100, but the rank of relative influence of variables are not

quite same. The top ranked variable influence from 200 trees model are: V39>V35>V46>V41>V6>V3. The top ranked variable influence from 2000 trees model are: V39>V35>V46>V41>V56>V29.

2) By summary comparison of predicted values from test dataset, the 200 tree model has average lower prediction accuracy (averaged predict 0 as 0.045, 1 as 0.3), but the 2000 tree model has a much better prediction accuracy (average predict 0 as 0.012, 1 as 0.54). Also, 200 tree model has a log loss value as 0.13 but 2000 tree has a log loss value as 0.09.
3) Thus, I concluded 2000 tree models are much better for predict the binary outcome bankruptcy. The best tree from this 2000 tree boosting model could be retrieved from 250~300 times iteration (See the figure below).



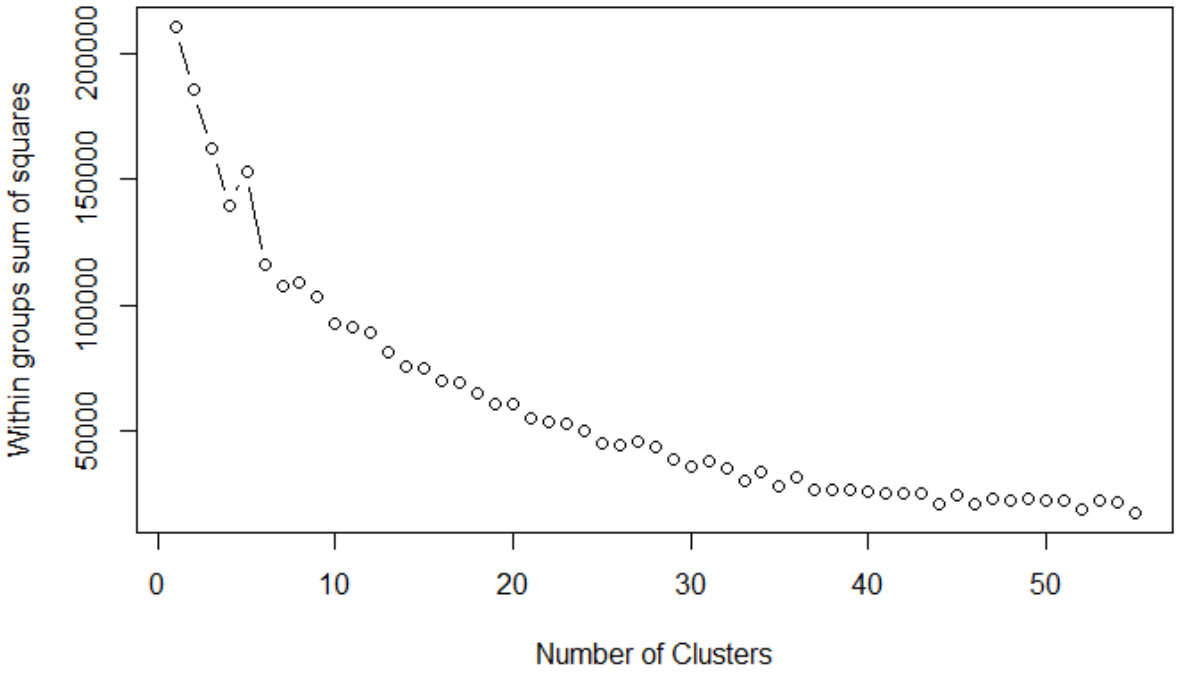
3 Support vector machines

1) Linear SVM: I fitted a simple linear SVM model with cost as 1, it gave 529 support vectors (273-0, and 256-1). Then I used the predicted value from this model to get the error rates of training dataset is 0.065 and the error rate of test dataset is 0.073. Next, I decided to tune the linear model with cost as 0.1, 1 and 10, and the tuning process indicates the best performance of linear SVM has error rate when the cost is lowest as 0.1, and tuning process did mild improvement over the course, with best performance has error rate as 0.066. The tuned best linear SVM predict that training set error rate is 0.066, and the test set error rate is 0.074. These numbers are close to the predicted errors from the original model.

2) Kernel SVM: I put the effort to tune a radial kernel SVM model. I used three cost options (0.1, 1 and 10), two gamma options (0.5 and 1). The tune process indicates the combination of 0.1 cost and 0.5 gamma gave the best performance, with best performance model has error rate as 0.068. Using the best tuned kernel SVM, the predicted error rate for training dataset is 0.042, and the error rate for testing data set is 0.072. Therefore, I concluded the tuned kernel SVM model with 0.1 cost and 0.5 gamma values gave best prediction over other parameter combination.

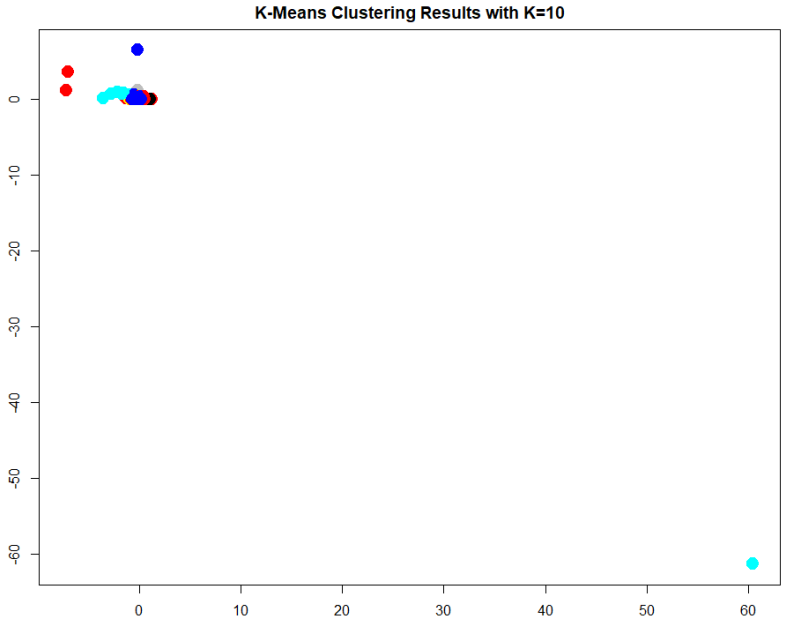
4 Clustering

I plotted the group sum of square and estimated the possible cluster number, I decide to fit 2 and 10 means clustering in following K-means clustering (Figure below).



1) K-means clustering:

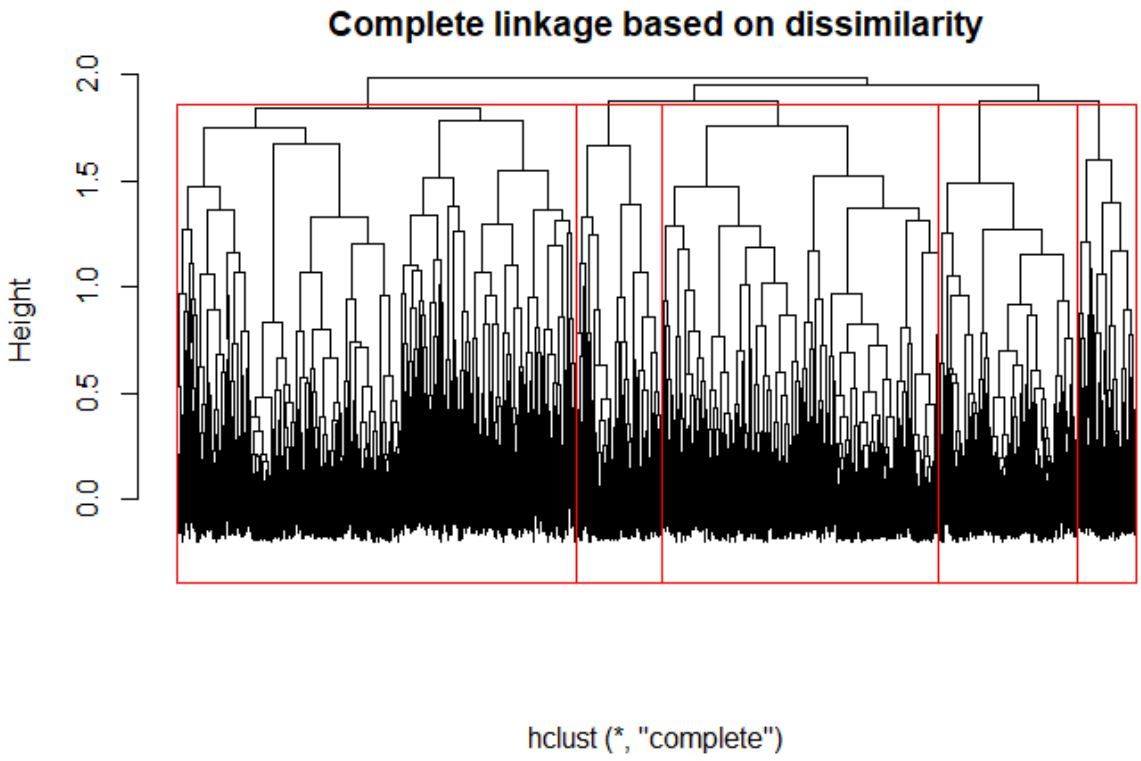
   From 2 mean and 10 mean clustering plot, I saw one cluster stand out from other clusters, I am not sure if the outcome is the only separated cluster, so I decide to remove outcome and try K-means cluster again, but I get the very similar plot.



2) Hierarchical clustering

Next, I explored hierarchical clustering approach. By Euclidean distance matrix, I got three type of dendrograms (complete, average and single), and these three plots are hard to visualize. Then I computed three dendrograms based on dissimilarity (1-Pearson correlation), these plots are easy to visualize.

3) I use cut function and partitioned the complete type of dendrogram into five parts (see below), and also compare these five parts and the outcome (see figure and table below). The outcome (1, yes, bankruptcy) predominantly located at 2nd part of hierarchical clustering dendrogram (193/273).



hc.clusters	0	1
1	1107	1
2	1405	193
3	555	0
4	334	10
5	175	58

I also compared 2-mean clusters and 5-cut endrogram (see table below). The table shows that the first cluster of k-means overlapped with majority of hierarchical clusters.

```
                hc.clusters
km.clusters    1    2    3    4    5
          1 1108 1595  555  344  228
          2    0    3    0    0    5
```

5 MDS

MDS uses dimension reduction techniques to identify a low-dimensional representation from high-dimensional data. MDS maps are constructed from information about proximity between any two point based on their distance. When the distance between any two points is small, MDS locates them near each other in the map; when the distance between two points is large, MDS locates far apart, and the final result of a MDS analysis is a statistical map. For cooperate' financial data, MDS map can provide intuitive interpretation in the setting of bankruptcy prediction.

Due to the large computation problem, I selected 100 records from bankruptcy training dataset. I then standardized the data, and obtained the distance matrix by cmdscale function. To plot the output, I extracted the two dimensional data points X1 and X2 and plotted a simple MDS map (see below).