Exploration and Analysis of data biopsy

Ruoyuan Li

1 Introduction

1.1 Data Description

The data set has been used in this analysis is called “biopsy”, where it can be found at “**MASS**” which is a **R** package. It is a breast cancer data was recoreded by Dr. William H. Wolberg who worked at University of Wisconsin Hospitals. The data contained information from biopsies of breast tumours for 699 patientes with 11 variables in total, one is “ID”, one is “class” which is response variable, and the other nine attributes are numerical variables which has been scaled of 1 to 10. This whole data is not complete, there are few missing value, so I take them off from the oringinal data, also I take “ID” off since it’s patients’ ID number which plays no role in this analysis. Then, the dataset left as a whole complete dataset which contain 683 observations with ten variables, nine of them are predictor variables, V1 is clump thickness, V2 is uniformity of cell size, V3 is uniformity of cell shape, V4 is marginal adhesion, V5 is single epithelial cell size, V6 is bare nuclei, V7 is bland chromatin, V8 is normal nucleoli. V9 is mitoses, again, V1 to V9 are scaled to range from 1 to 10. And the other one is binary response variable called “class” with 2 levels “benign” and “malignant”, 444 observations are “benign” class, the rest 239 observations are “maglinant” class.

The Figure 1 is the correlation plot of “biopsy” data, the orange colour represents class “bengin”, and the blue colour represents class “malignant”. As we can see, there is no clearly lienar relationship between those predictor variables. However, we can see the “benign” class are most ranges from 0 to 5, and “malignant” class ranges vary 1 to 10. There is no strong postive or negative correlations based on correlated values shows on upper right plot matrix. Also, Table 1 is basic statistical summarizes of biopsy data, as I mentioend before, V1 to V9

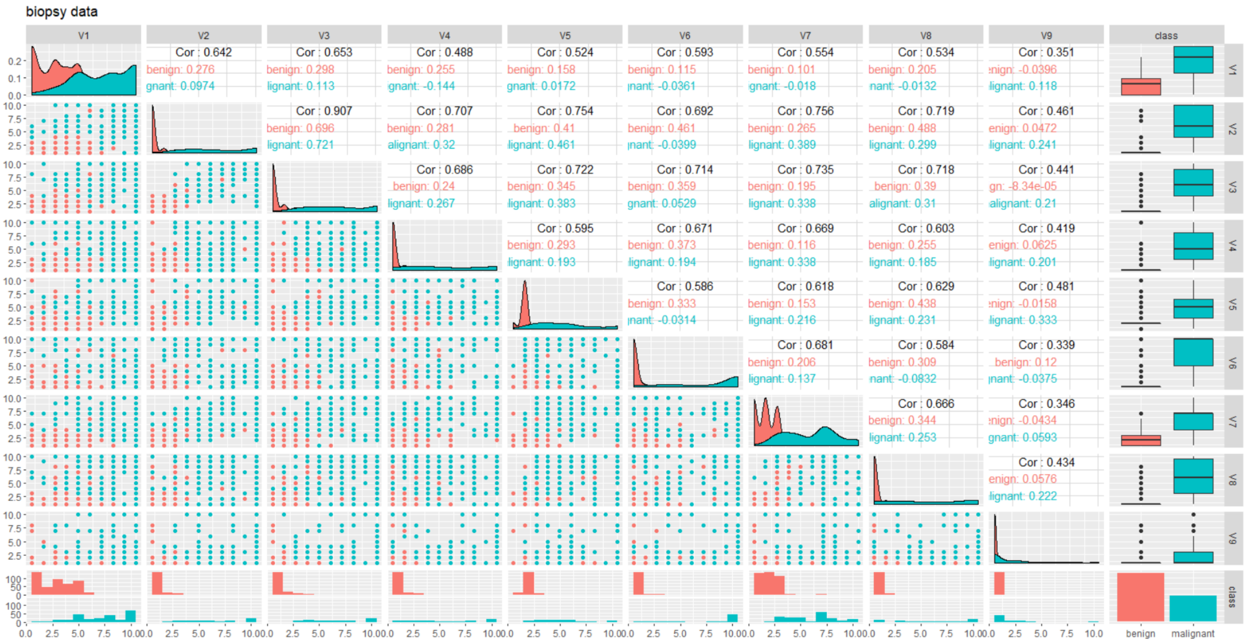

Figure 1: correlation plot of biopsy data

are scaled to 1 to 10, so the minimum is 1 and maximum is 10 for all numberical variables, and one factor variable “class”.

1.2 Questions Addressed

In this report, I would like to explore and analysis the biopsy data by doing few classification method since it may helpful for diagonasis the breast cancer or other breast realted tumours on early disease period for new patients. The goal is find the best classification method with best prediction performance.

The analysis will following with description of each classification method, and explain the choice of parameters. Then, choose the best model with suitable parameters, run 10 different random 75/25 labelled/unlabelled splits for generalized the classification results. Based on all the results from all those classification method, there is a discussion part for comparing those method.

Table 1: Summary Statistics of data biopsy

variable	type :	numeric	n=683				
variable	mean	sd	p0	p25	median	p75	p100
V1	4.44	2.82	1	2	4	6	10
V2	3.15	3.07	1	1	1	5	10
V3	3.22	2.99	1	1	1	5	10
V4	2.83	2.86	1	1	1	4	10
V5	3.23	2.22	1	2	2	4	10
V6	3.54	3.64	1	1	1	6	10
V7	3.45	2.45	1	2	3	5	10
V8	2.87	3.05	1	1	1	4	10
V9	1.6	1.73	1	1	1	1	10
variable	type:	factor	n=683				
variable	n_unique			ordered			
class	2 levels	ben:444	mal:239	FALSE			

2 Classification Analysis

The main techniques is classification method, such like classification tree, bagging, random foreest, boosting, and model based classification such as using **pgmm** package, **teigen** package. The following part will discuss the detail of each method and the parameters setting. As I mentioned before, first discuss theory, then show steps for choosing parammeters and results in **R** for each method. Next run random 75/25 labelled/unlabelled splits for ten times after findding the suitable paramters. Some method can do corss validation for parameter, some can not, however, the detail is shown as following part for each method. The overall comparsion for 10 times randomly splitting is showing at next section.

2.1 Classifcation Tree

Classifiction tree defined as it starts at root and spliting data recursively based on best splitter. Based on Hastie et al. (2009), a node m represents a region R_m with N_m observations, the proportion of observations from class g in node m is defined as

$$\hat{p}_{mg} = \frac{1}{N_m} \sum_{x_i \in R_m} \mathbb{I}(y_i = g)$$

so that all observations under node m are all classified into majority class which means the largest proportion of the observations.

For the steps in **R**, frist randomly sample 75% of observations as training set, the rest 25% observations as test set. Then call function **tree** under package **tree** in **R**, sepcify the formula as $class \sim V1 + \cdots + V9$, method is "class" which means classification, running it for training data set. This tree ends with 9 terminal nodes, the misclassification error rate is 2.15% which is very good, and the residual mean deviance is 0.1023. The actual tree is shown as Figure 2, also notice only 5 out of 9 predictior variables are using for constructing the tree, namely V3, V1, V6, V2, and V5. Next based on tree which builds by traning data, using the tree predicting the test data, and make a table to compared predict class with actual class of test data which is Table 2. The result table is very good, with only 6 misclassified observations out of 171. Also calculting the performance by **e1071** package with fucntion **classAgreement**, the table has 3.5% misclassified rate and ARI is 0.862 which are quite good. Next is repeating this process for nine more times, save all the values which calcualted by **classAgreement**. Those results will be shown later.

Table 2: prediction vs actual class by classification tree

	benign	malignant
benign	112	4
malignant	2	53

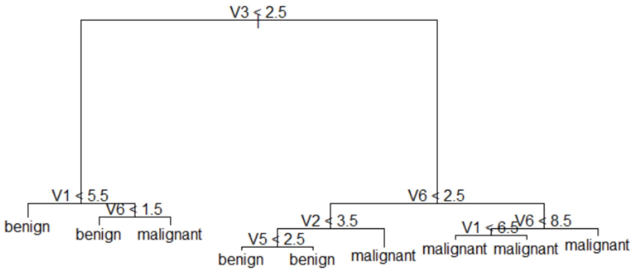

Figure 2: classification tree for biopsy data

2.2 Bagging & Random Forrest & Boosting

Bagging defined as first use bootstrapping to resample the training set, and fit a learning method to each bootstrap sample, finally averaging the resulting predictions. In mathematical expression can be defined as:

$$\hat{f}_{bag}(x) = \frac{1}{M}\sum_{m}^{M}\hat{f}^{m}(x)$$

where M represents generates M bootstrap ensembles from the training set, and $\hat{f}^{m}$ is prediction trained on the mth bootstrap ensemble.

There exist the situation as some observations are used more than oce, but some observations will be left altogether when we creat bootstrap ensemble, the left observations are called “out-of-bag”. The mean square error of out-of-bag can be calculted as:

$$MSE_{obb} = \frac{1}{M}\sum_{m=1}^{M}(\hat{f}^{m}(x) - \hat{f}_{bag}(x)$$

Random forrest is extension version of bagging, random forrest decorrelates the M trees. Unlike bagging using all predictor variables avaliable for each split, a random foreest only take M predictors for each each split which M is randomly sampled. Compared to bagging, random foreest preventing the situation such as same predictor variable dominates the tree. Usually M equals to $\sqrt{p}$ where p is the number of predictor variables.

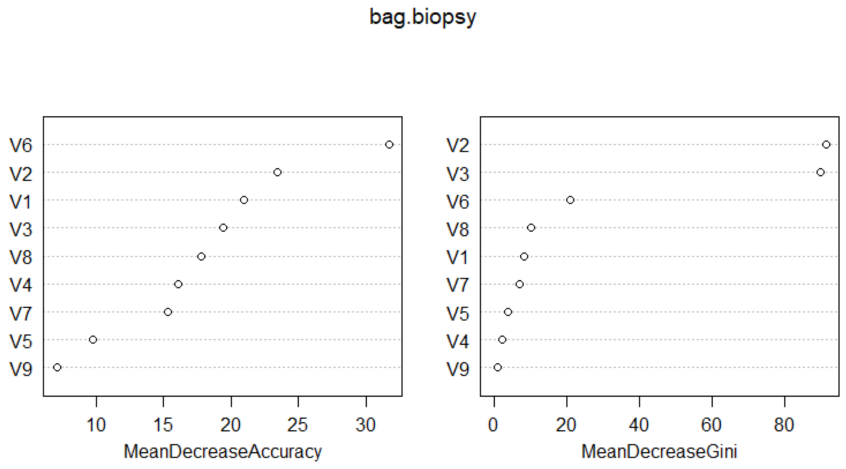

Figure 3: Variables importance plot from bagging

For boosting, one similarity is it can used beyond classification tree like bagging, one disimilarity is boosting grows trees sequentially instead of independtly growing like bagging. The current residual is response for each tree.

For applying those three methods in **R**, the two package **randomforrest** and **gbm** are using. First running bagging and random forrest, at the begining sample 75% of the observations as training set, the rest 25% of observations is test set. As for bagging, using all predictor variables for building the tree, so I speicify the paramter "mtry" equals to 9 since biopsy data has nine predictor varialbles, the number of trees is 500 by defualt, same formula is using which is $class \sim V1 + \cdots + V9$, the data is training data. Then get the results of bagging, OBB rate is 2.93% with 500 trees and 9 variables tried at aech split. Also checking variables importance plot as hown at Figure 3, for MeanDecreaseAccuracy plot we can see V6 and V2 has higher accuracy value which means they contribute more than other, they are ranked higher in usefulness than other variables for correctly classifying data. And for MeanDecreaseGini plot we can see V2 and V3 are ranked higher in usefulness for correctly classifying data than other variables.

After examining bagging results from training data, predicting the test data based on bagging object and make a table with actual class of test data and predict class. The result's table is Table 4, the misclassifcation rate is 4.67% and the ARI is 0.818. This result is a little worse

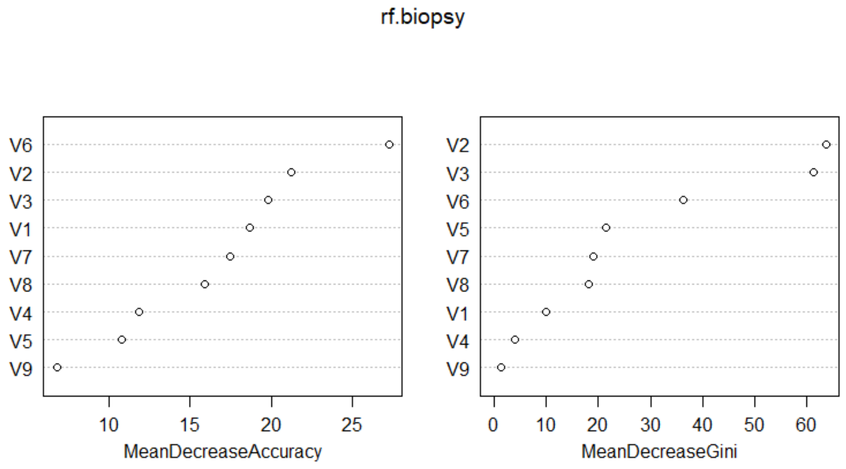

Figure 4: Variables importance plot from random forrest

than classification tree from before, but it may casues by sampling process, so again running the bagging for nine more times, save all missclassification rate and ARI for showing later.

As for random forrest, first I go through general parameter selecting which means change parameters manually for selecting purpose, then cross validation can be used for random forrest. All the setting are the same as bagging, exclude changes "mtry" equals to 5 and 3 for trying, then examinning their OOB rate. Then for 5 variables tired at each split with OOB rate is 2.54%, and for 3 variables tired at each split with OOB rate is 1.95%. So far 3 variables tired at each split with default 500 trees has better OOB rate. Then using cross validation method, there is **tune.randomforrest** function in package **e1071**. Specifying "mtry" ranges from 1 to 9 which represents nine varialbes, and specifying the trees from 100 trees to 500 trees, setting 5 cross within tune.control option. Then running random forest validation and ending with best random forrest model's parameters setting are "mtry" equals to 1, and 200 trees. Based on best parameters selecting by cross validation, then use it in regular random forrest, also compared OOB rate as before, it ends with OOB rate equals to 2.93% which is higher than random forrest models' from before. So it may not a good choice to choose parameters setting suggests by tune process. Finally, the best random forrest result is bulding by 3 variables tired at each split and 500 trees. Then examine the varaibles' imprtance plots as well, it shown as Figure 4.

Similar procedures are using for random forrest, using random forrest obejct from training data to predict test data, and make a table with predict class and actual class from test data. Te result's table is Table 4, it has misclassification rate equals to 3.5% and ARI equals to 0.862. We can see random forrest prvides better results than bagging so far. However, repeat this process for nine more times to generalized the results, the detail will be shown later as well.

Next move to boosting using **gbm** package in R, same as before, using training data for building boosting model, and predict based on test data. Also notice biopsy data has binary response, so in **gbm** function specifying distribution to be "bernoulli", the number of trees using for building boosting is 5000 trees for each cases. There are few different parameters setting for running **gbm** function, the first one running gbm for training data with interaction depth equals to 4, λ equals to 0.001, and ntree is 5000 for prediction; the second one is changing λ to 0.01 and keeping interaction depth as 4 still, and ntree for prediction is choosing by **gbm.perf** which is 427 trees in this case, also consider if 427 trees is under estimate, so try 727 trees as well (300 more trees than suggested by **gbm.perf**); the thrid one is reducing λ to 0.01 and reducing interaction depth to be 1, the tree for prediction also suggesting by **gbm.perm** which is 544 in this case, also try 844 trees in case of under estimation. The Table 3 is comparison table for all differernt parameters setting for boosting, ARI are pretty similar for each cases, it may causes by binary response varialbes with distribution to be bernoulli.

Table 3: Bossting's ARI table from different parameters setting

interaction depth	4	4	4	1	1
n.tree (prediction)	5000	427	727	544	844
shrinkage (λ)	0.001	0.01	0.01	0.01	0.01
ARI	0.8617	0.8617	0.8617	0.8619	0.8398

However, consider the calculting time, the better one may be the one with 544 trees and shrinkage equals to 0.01. Next review the detail for this cases. The gbm function for this case defines as: $gbm(class\ ., data = biopsy_1[train,], distribution = "bernoulli", n.trees = 5000, interaction.depth = 1, shrinkage = 0.01)$. According to summary and varaibl relative

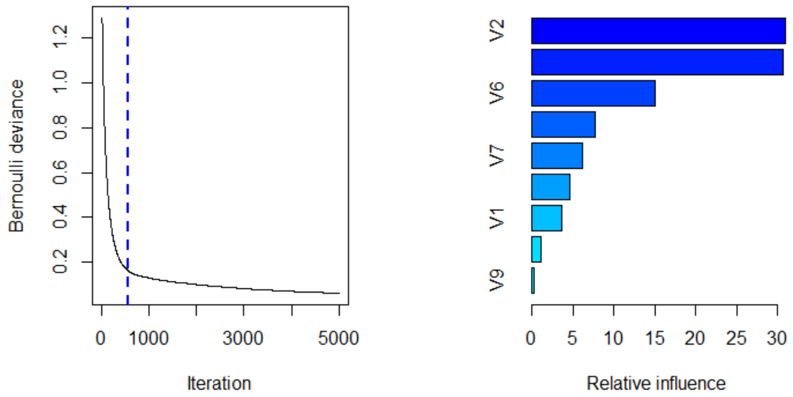

Figure 5: ntree selecting by gbm.perm plot and variable relative influence plot

plots shown as Figure 5, V2 and V3 contributes more than other varaibles for correctly classifying data. Also predicting test data, the trees using prediction is 544 trees. And make a table with actual and predict class, the table is shown as Table 4 as well. Repsting this boosting for nine more times, save ARI and misclassification rate, those results is showing later.

For those 3 bootstraping method, bagging, random forrest and boosting, the table showing similar results. Even including classification tree result, they are all similar to each other.

Table 4: Result table from Bagging, Random Forrest and Boosting

bagging	B	M	RF	B	M	Boosting	B	M
B	111	5	B	112	4	B	112	4
M	3	52	M	2	53	M	2	53
ARI	0.818		ARI	0.862		ARI	0.862	

2.3 Model based Classification: teigen package in R

McNicholas and Murphy (2008) develop a family of eight parsimonious Gaussian mixture models (PGMMs) for clustering, and nowdays non Gaussian mixture models attarcts more attention than before, so *t*EIGEN family (Andrews and McNicholas 2012; Andrews, Wickins,

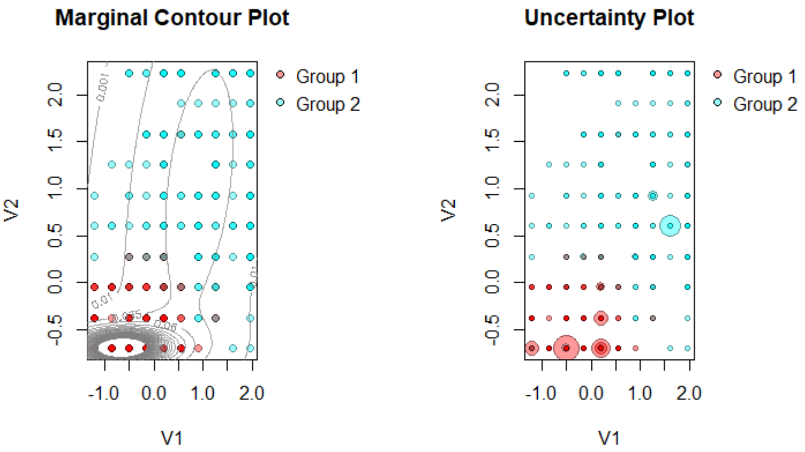

Figure 6: Marginal Contour Plot and Uncertainty Plot from teigen

Boers, and McNicholas 2018) comes up with more models avaliable, it is model based clustering and classification with the multivariate t distribution, whose density is

$$f(\mathbf{x}|\vartheta) = \sum_{g=1}^{G} \pi_g f_t(\mathbf{x}|\mu_g, \Sigma_g, \upsilon_g) = \sum_{g=1}^{G} \pi_g \frac{\Gamma(\frac{\upsilon_g+p}{2})|\Sigma_g|^{-\frac{1}{2}}}{(\pi\upsilon_g)^{\frac{p}{2}}\Gamma(\frac{\upsilon_g}{2})[1+\frac{\delta(x,\mu_g|\Sigma_g)}{\upsilon_g}]^{\frac{\upsilon_g+p}{2}}}$$

where μ_g is mean vextor, Σ_g is scaled matrix, and υ_g is the degree of freedom. The eigen-decomposition affect Σ_g which is scaled matrix, later the EM algorithm is imposed, EM algorithm represents 2 steps, E step which is calculting the expected value of the complete-data log-likelihood, M step is the parameters are maximized according to the complete-data log likelihood. There are 28 models list in teigen package (Andrews, Wickins, Boers, and McNicholas 2018), the list from “CIIC” to “UUUU”, ‘C’ represents constrained, ‘U’ represents unconstrained, and ‘I’ respresents the identity matrix. ‘*’ represents the models being introduced. BIC and integrated completed likelihood (ICL, Biernacki, Celeux, and Govaert 2000) are using for model selection in teigen. Next move to process in R code.

First scale all predictor varialbes V1 to V9, and specified class to be “class” from biopsy, then randomly sample 25% of class to be NA, so the left 75% class is known class. And “Gs” equals 1:2 which defines as indicating the number of groups to fit, “init” as “uniform” which is using for classification only, run teigen in paralle, “known” is known class from before which is 75% of total observations. Then it ends with best model suugestion is “UUCU” with

“G=2” and BIC is -4844.64. As shown at Figure 6, the left is a bivariate marginal contour plot,it specifys the desired variates and the resolution of the contours. The second plot is an uncertainty plot, larger dots represent large uncentainty from classification, so there are more uncertanity from group 1 than group 2.

Then fitting teigen again with “model” specified to be “UUCU”, keeping other parameters to be same as before.

And making a table with classification results from teigen object with actual class of biopsy data. Finally the table is shown as Table 5 and ARI is better than all the trees' method from previous section. Also the good results may come randomly, so run this process nine more times, and the overall results is shown at Discussion section.

Table 5: Results' table from teigen and QDA

teigen	1	2	QDA	benign	maglinant
bengin	101	2	benign	108	2
maglinant	0	67	maglinant	2	58
ARI	0.953		ARI	0.907	

2.4 QDA (quadratic Discriminant Analysis) with PCA

The QDA represents quadratic discriminant analysis, it is a classification method. It supposes if there are G class, then it assumes:

$$p(\mathbf{x}|class\ g) = (2\pi)^{-\frac{p}{2}}|\Sigma_g|^{-\frac{p}{2}} \exp\{-\frac{1}{2}(\mathbf{x}-\mu_g)'\Sigma_g^{-1}(\mathbf{x}-\mu_g)\}$$

and let π_g to be the probability of an observations from class g, π_g defines as

$$P[\ class\ g|\ x\ ] = \frac{\pi_g p(x|classg)}{\sum_h^G \pi_h p(x|classh)}$$

if thre are class g and h, the observaion x is classified into class g if

$$P[\, class\, g|\, x\,] > P[\, class\, h|\, x\,]$$

Also note the difference between LDA (linear discriminnat analysis) and QDA (quadratic discriminant analysis) is LDA has assumption which states $\Sigma_g = \Sigma$ for $g = 1, \ldots, G$.

Principle components analysis (PCA) uses for finding a number of uncorrelated variables which explain enough variance in the observed data. The first principal component is the direction of most variation (in the data). And for $r > 1$, the rth principal component is the direction of most variation (in the data) conditonal on it being orthogonal to the first $r - 1$ principal components. Next move to **R** for details.

Frist doing the PCA for V1 to V9, calling **prcomp** function and set scale to be TRUE. Then checking the summary of PCA, in genenral choosing the numebr of principla component if their standard deviation around 90%, then PC1 and PC2 would be two good choices.

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6
## Standard deviation     2.4289 0.88088 0.73434 0.67796 0.61667 0.54943
## Proportion of Variance 0.6555 0.08622 0.05992 0.05107 0.04225 0.03354
## Cumulative Proportion  0.6555 0.74172 0.80163 0.85270 0.89496 0.92850
##                            PC7     PC8     PC9
## Standard deviation     0.54259 0.51062 0.29729
## Proportion of Variance 0.03271 0.02897 0.00982
## Cumulative Proportion  0.96121 0.99018 1.00000
```

Next take a look of correlation plot of PC1 and PC2, there is distinguish difference between each class, it may indicates it may be good choice to do QDA classification based on PC1 and PC2. Then as usual, divide data into 75% traininig set which is lablled observations and 25% test set which is unlablled observations, using traing set to build the model, and the test set using for checking the efficacy of the model. Then the first run results shown as Table 5, the AIR is slightly better than all those trees' method from before. Also, doing this process for nine more times and results shown at Section Discussion.

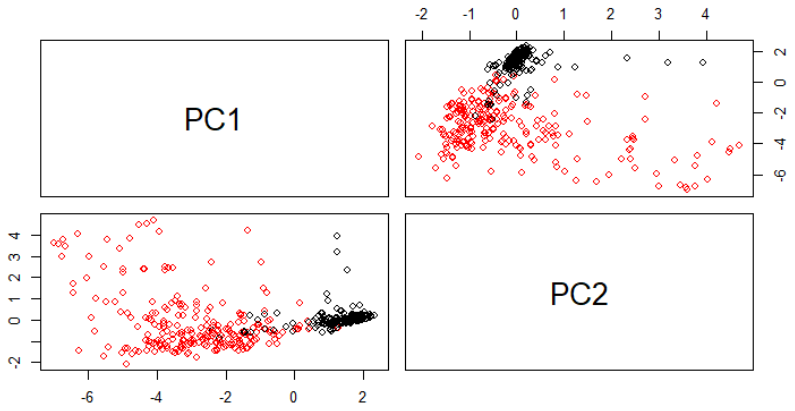

Figure 7: correlation plot for PC1 and PC2

3 Discussion

So far I tried 6 method in total, and in this part will discuss model efficacy of each method by recording all the ten times labbled/unlabbled random splits. The detail for each methods have been discussed before, next move to