Assignment:
Multivariate Data Analysis in R

Applied Multivariate Statistical Analysis

Prof. Eddie Schrevens

Hamed Borhani – r0438210

Master of Bioinformatics

2015-2016

KU Leuven
Faculty of Bioscience Engineering

1.Introduction

1.1.Overview

In this report, multivariate data analysis techniques in R are exploited to solve a problem.
First of all, a general description about the problem and the used dataset is given.
The R code, the dataset and other additional files in this assignment are available in the github repository: *https://github.com/hamed2005*

1.2.Problem description

microRNAs (or miRNA) are small non-coding RNA molecules (around 22 nucleotides) that function in post-transcriptional regulation process in plants and animals.
mRNAs (messenger RNA) are the molecules that carry the codes from DNA in the nucleus to the cytoplasm, where they would be translated into proteins.
A miRNA would attach to it's target mRNA(s) and forms a miRNA::mRNA duplex and block it's further translation into proteins. Predicting miRNA targets is very important, because they play a vital role in many biological processes including cell proliferation, cell death, hematopoiesis and oncogenesis. So, they can help researchers in finding out the real causes of diseases like lymphoma, leukemia, cancers and many cardiac problems where miRNA:mRNA pairing is found to play a crucial role.
This process depends highly on the sequence identity between the miRNA and it's target, but a lot of other factors also play a role since microRNA could target a mRNA even without 100% sequence identity.
In this report it is tried to gain an insight into the associated variables with miRNA and it's target and investigate the data in order to find a way to predict if a miRNA can target a specfic mRNA or not.

1.3.Dataset description

The dataset is obtained from the paper "*HomoTarget: A new algorithm for prediction of microRNA targets in Homo sapiens*"[1], in which they combined PCA and Pattern Recognition Neural Network (PRNN) to predict the miRNA targets.
This dataset totally has 425 observations in 2 classes (112 "positive" and 313 "negative"). Each observation has 15 variables, a miRNA sequence (1st) and the target sequence (2nd) are omitted in the analysis. The variables used in the analysis are as follows:

#	**Variable**	**Description**
3	Total Score	Obtained by the sum of pair scores. Match +5, G:U +1, Mismatch −3, Gap −1
4	Seed Score	Obtained by the sum of pair scores in the seed region
5	WC Pairs	Number of WC pairs in the duplex
6	Wobble Pairs	Number of wobble pairs in the duplex
7	Mismatches	Number of mismatches in duplex
8	Number-bulges	Number of bulges in the duplex
9	A proportion	Proportion of "A" in the duplex
10	C proportion	Proportion of "C" in the duplex
11	G proportion	Proportion of "G" in the duplex

12	U proportion	Proportion of “U” in the duplex
13	A:U proportion	Proportion of A:U matches in the duplex
14	Minimum Free Energy	Calculated using RNAfold (from Vienna RNA package) for a duplex formed by the miRNA and its target.

All variables are between 0 and 1 except the scores which are between -1 and 1 and the minimum free energy that ranges in [-25,0].
The dataset is in the spreadsheet XSLX format and each class is provided in a separate sheet and There are no missing values. For the ease of use, the XSLX file was converted to two CSV files (one for positive and one for negative) using the *ssconvert* command line utility.

2.Proposed multivariate analysis and the motivation

The main problem here is a discrimination problem. Therefore discriminant analysis techniques, namely “Linear Discriminant Analysis” and “Decision Tree” were performed on the dataset. Before doing any analysis, outliers were detected using LOF (Local Outlier Factor) algorithm and visualized in the Biplot (number of observations are shown). Those which were obviously outliers, were removed from the dataset. Biplot also shows correlations between variables, therefore first 9 principal components were fed into LDA instead of the variables themselves, which resulted in 92% accuracy.
Tree based modeling was also used (with all available variables) to build a model which reached 92% accuracy with only 3 terminal nodes and using just 2 variables which is a significant dimension reduction.

3.R Code

```
library(plotrix)    ## for biplot
library(rpart)      ## for Tree-based Modeling
library(MASS)       ## for linear discriminant analysis
library(DMwR)      ## for outlier detection

## setting the working directory
setwd("/home/hamed/KUL/Multivar/Project/HomoTarget")
## reading the CSV files for positive and negative instances
colnames <- c("mirnaSeq","targetSeq", "totalScore", "seedScore", "WCPairs",
"WobblePairs", "mismatches", "NumberBulges", "A", "C", "G", "U", "AU",
"minFreeEnergy")
pos <- read.csv("dataset.csv.p", header = FALSE, col.names = colnames)
neg <- read.csv("dataset.csv.n", header = FALSE, col.names = colnames)
## appending two dataframes into one
mirnaDF <- rbind(pos,neg)
## omitting the sequence columns
mirna <- mirnaDF[,3:14]
## overview
summary(mirna)
multi.hist(mirna)
## pairs plot
pairs(mirna, pch=20, col="#383838")

#removing outliers detected using “lofactor” from “DmwR” package
mirna <- mirna[-103,]
mirna <- mirna[-94,]

## centering and normalizing the data (unit variance)
mirna.c <- scale(mirna, scale = F, center = T)
```

```
mirna.mean <- attr(mirna.c, "scaled:center")
mirna <- mirna.c[,]
mirna.n <- scale(mirna, scale = T, center = T)

## variance-covariance and correlation matrix extraction
cov(mirna)

### Biplot
x <- mirna
xm<-apply(x,2,mean)
y<-sweep(x,2,xm)
ss<-(t(y)%*%as.matrix(y))
s<-ss/(nrow(x)-1)
d<-(diag(ss))^(-1/2)
e<-diag(d,nrow=ncol(x),ncol=ncol(x))
z<-as.matrix(y)%*%e
r<-t(z)%*%z
q<-svd(z)
gfd<-((q$d[1])+(q$d[2]))/sum(q$d)
gfz<-(((q$d[1])^2)+((q$d[2])^2))/sum((q$d)^2)
gfr<-(((q$d[1])^4)+((q$d[2])^4))/sum((q$d)^4)
l<-diag(q$d,nrow=ncol(x),ncol=ncol(x))
R.B<-q$u        #scores matrix
C.B<-q$v%*%l    #loadings

#possibility to stretch scores by a scale factor
scalefactor<-3.5
R.B<-q$u *scalefactor

par(mar=c(4,4,4,4),pty='s',oma=c(5,0,0,0),font=2)
plot(R.B[ ,1],R.B[ ,2],axes=F,xlim=c(-1,1),ylim=c(-1,1),xlab=' ',ylab=' ',cex=.8)
mtext('First component',side=1,line=3,cex=.8)
mtext('Second component',side=2,line=3,cex=.8)
axis(1,at=c(-1,-.8,-.6,-.4,-.2,0,.2,.4,.6,.8,1),cex=.8)
axis(2,at=c(-1,-.8,-.6,-.4,-.2,0,.2,.4,.6,.8,1),cex=.8)
box( )
points(R.B[,1],R.B[,2],pch=".")
points(C.B[,1],C.B[,2],pch=".")
text(C.B[,1]-.05,C.B[,2]+.05,as.character(dimnames(x)[[2]]),cex=0.8)
## drawing the arrows
for (i in seq(1,nrow(C.B),by=1))
        arrows(0,0,C.B[i,1],C.B[i,2])
#Draw circle unit
draw.circle(0,0,1,border='black')
mtext('PCA Biplot',side=1,outer=T,cex=1,line=3)

par(mar=c(5,4,4,2) + 0.1)    ## to reset margin settings
par(oma=c(3,3,3,3))

### PCA
mirna.pca <- princomp(mirna, cor = T)    ##using correlation matrix
screeplot(mirna.pca, type="lines")
mirna.pca$loadings
summary(mirna.pca)

### Linear Discriminant Analysis
## LDA using first 9 PCs with cross-validation
##adding class variables
pos<-cbind(pos,'positive')
colnames(pos)[15]<-"class"
neg<-cbind(neg,'negative')
colnames(neg)[15]<-"class"
##merging pos and neg to build the full dataset
```

```
mirna <- rbind(pos,neg)[,3:15]
mirna$class <- as.factor(mirna$class)
##removing the outliers
mirna <- mirna[-103,]
mirna <- mirna[-94,]
mirna.lda <- lda(mirna$class~+mirna.pca$scores[,1:9], CV=T)
head(mirna.lda$posterior)
table(mirna$class, mirna.lda$class, dnn = c("From", "classified into"))

### Tree based modelling -tree construction using rpart
mirna.rDTree <- rpart(class ~ .,
                      data = mirna , method="class",
                      parms=list(prior=c(0.5, 0.5)))
par(xpd=NA)    ## to show the labels completely in output
##results of cross validation
plotcp(mirna.rDTree)
##pruning
mirna.rDTree.pruned <- prune(mirna.rDTree, cp = 0.062)
plot(mirna.rDTree.pruned, uniform = F, main = "Pruned Decision Tree")
text(mirna.rDTree.pruned, use.n = T, all = F, cex = 0.7)
print(mirna.rDTree.pruned)
##misclassification error and CV for pruned tree
out <- predict(mirna.rDTree.pruned) # predict probabilities
pred.response <- colnames(out)[max.col(out, ties.method = c("random"))] # predict
response
table(mirna$class, pred.response)
```

4.R output

Additional outputs are available in the github repository.

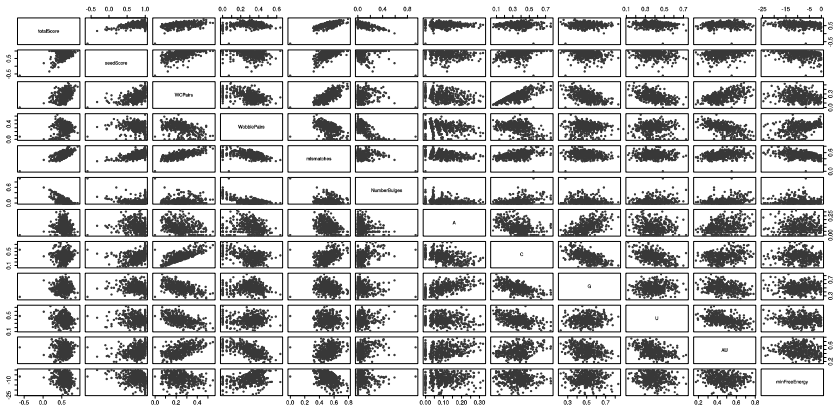

Figure 1: Pairwise Scatterplots

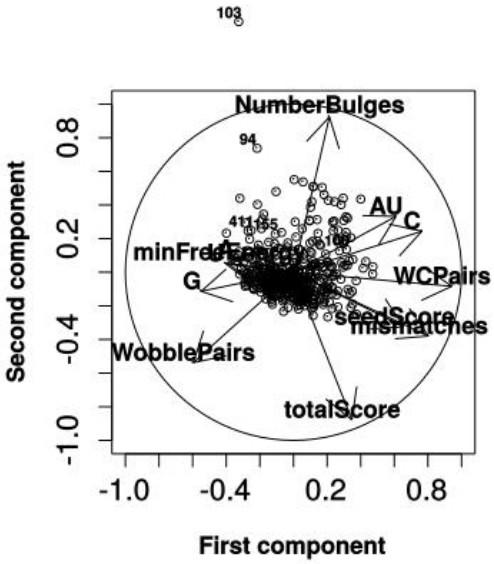

Figure 2: Biplot with labeled outliers

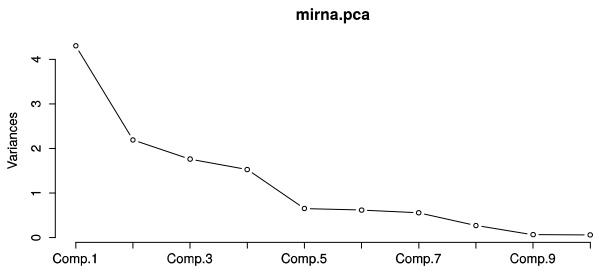

Figure 3: PCA scree plot

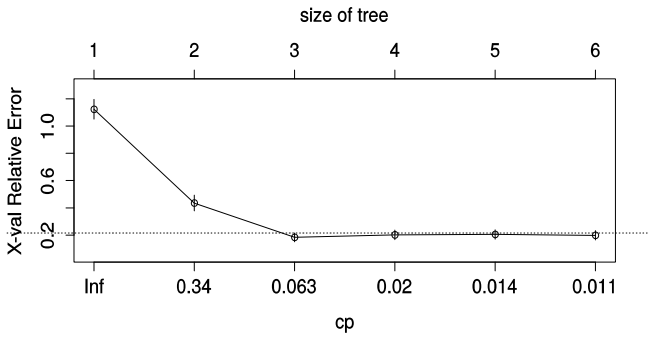

Figure 4: cross-validation result for TBM

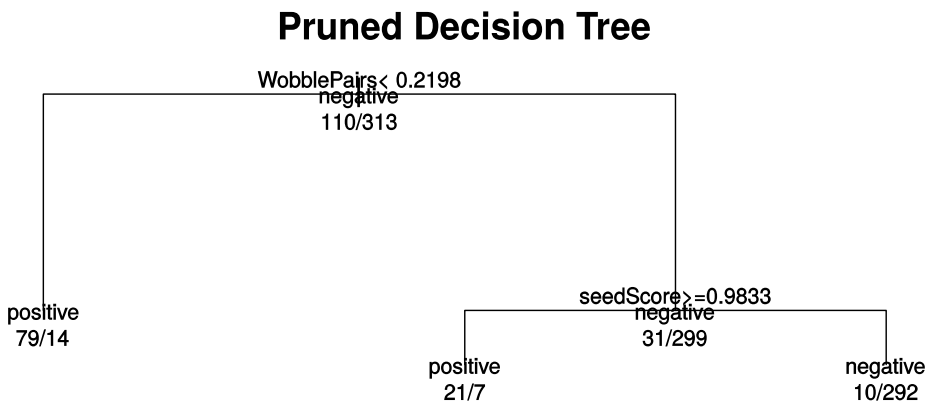

Figure 5: Final Pruned Tree

Characteristics of nodes in the final pruned tree

```
1) root 423 211.500000 negative (0.50000000 0.50000000)
  2) WobblePairs< 0.2198065 93   9.460064 positive (0.94137130 0.05862870) *
  3) WobblePairs>=0.2198065 330  59.604550 negative (0.22780739 0.77219261)
    6) seedScore>=0.9833335 28   4.730032 positive (0.89513823 0.10486177) *
    7) seedScore< 0.9833335 302  19.227270 negative (0.08879433 0.91120567) *
```

5.Interpretation and Discussion

Compared to the complex models used in the corresponding paper, simple and powerful discriminators made in this assignment are doing a good job. The decision tree reached 92% accuracy (based on the cross-validation result) using only 2 variables and having just 3 terminal nodes. This is also considered a very good dimension reduction compared to PCA which suggested 9 PCs out of 12. The only problem with the decision tree is it's high false positive rate. The LDA model in the other hand also has a problem which is high false negative rate. A solution could be to combine these two discriminators (e.g. using stacking) to further lower the errors.

Although it's indeed a discrimination problem, it is useful to perform a cluster analysis on the dataset. The main goal for doing so is to see whether we can relate each cluster to a different cell state such as normal or under stress conditions or expression in different tissues. By looking at profile plot (and andrwes plot) and also star plot, existence of grouping structures is supported. Cluster analysis was also done using hierarchical cluster analysis suggesting existence of 3 clusters.
However to continue, further expert knowledge and probably more meta-data is needed.

Factor analysis is a also a possibility to group the highly correlated variables.

7.Reference

[1] H. Ahmadi, et al. (2013). Homotarget: a new algorithm for prediction of microRNA targets in Homo sapiens. *Genomics, 101 (2013)* 94–100.