Bioinformatics Lab 5, Group 3

Roshni Sundaramurthy, Prudhvi Peddmallu, Jiawei Wu, Zijie Feng

11 December 2018

Rule-based classification and visualization

Task 1: Loading the required packages

```
#install.packages("devtools")
library(devtools)
#install_github("mategarb/R.ROSETTA")
library(R.ROSETTA)
```

Task 2: Loading the autcon dataset

```
autism_df<-autcon
#View(autism_df)
confusion_matrix <- table(autism_df$decision)
confusion_matrix
```

```
##
##  autism control
##      82      64
```

Analysis:

Description of dataset autcon

It is a sample dataset of gene expression values. The objects are divided into two decision classes: male children with autism and healthy ones. The features are represented by genes. There seems 146 children datas and 35 genes. The class variable is "Decision". This variable differentiates the children whether they are affected by autism or not.

Number of features: 35 Number of objects: 146

The confusion matrix is created using table function. The result seems to have two decision classes

1. Male children with autism (82)
2. Healthy male children (64)

So, we can conclude that the distribution of objects is not balanced.

Task 3: Using rosetta() on the default parameters

```
#autconDefault <- rosetta(autism_df)
#save(autconDefault, file="autconDefault.Rdata")
load("autconDefault.Rdata")

# Rule table information
tab_info <- autconDefault$main
```

```
# quality statistics of the model
model_stats <- autconDefault$quality

# Significant rules
library(dplyr)
sig_rules <- tab_info %>% filter(PVAL<0.05)
cat(paste("Number of obtained rules:", count(sig_rules)))
```

```
## Number of obtained rules: 185
```

Analysis:

a) **Cross validation:**

Cross-validation (CV) is a technique to assess the generalizability of a model to unseen data. This technique relies on assumptions that may not be satisfied when studying genomics datasets. The data is partitioned into equally sized subsets. k-fold cross-validation is the most commonly used technique for model assessment.

The argument "cvNum" in rosetta is a numeric value of the cross-validation number. The default is **10**.

b) **Default reduction method:**

The default reduction method is *Johnson* reducer method, or Johnson-Lindenstrauss theorem. It is used to map original data from a high dimension space into a low dimension space at an enough small cost.

c) **Default method of discretization:**

The default method of discretization is *EqualFrequency* method. It is a single-variable unsupervised-learning equal-frequency algorithm. It distributes nearly equal number of objects in different intervals according to the histogram of each variable. The default number of discretization bins calculated is *3*.

d) **Accuracy of the model:**

According to the `model_stats`, the accuracy of our model is shown as follow:

Accuracy	Value
Mean	0.821818
Median	0.8
Std	0.083158
Min	0.733333
Max	1

e) Rules:

Totally 185 rules have been obtained. The class **control** gets more significant rules.

```
cat(paste("Top 3 significant rules:"))
```

```
## Top 3 significant rules:
```

```
head(tab_info,3)
```

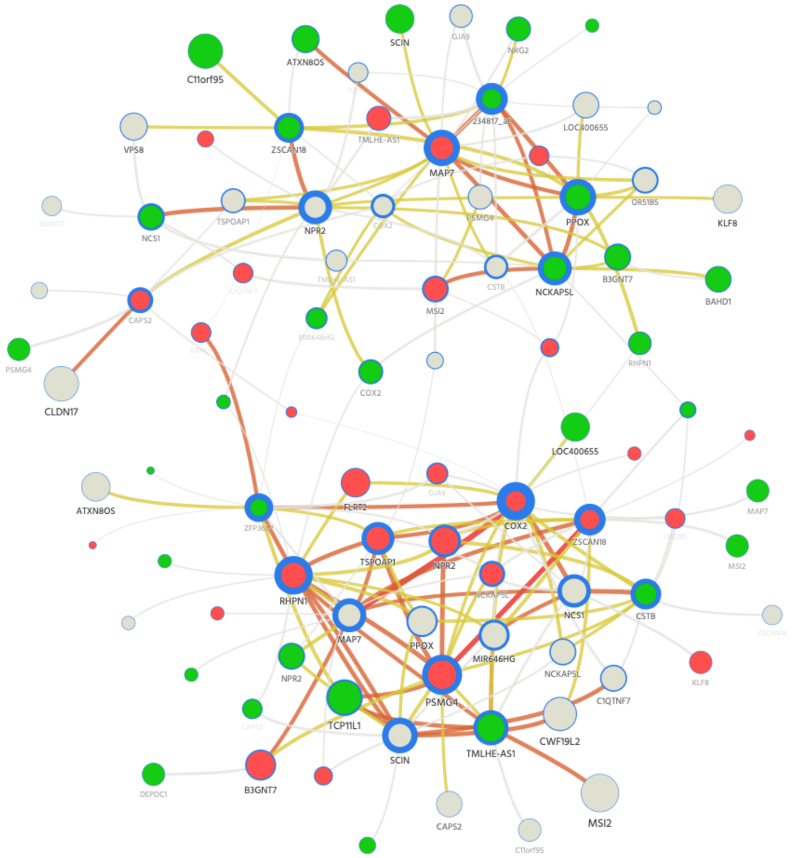
```
##            FEATURES DECISION               CUTS_COND DISC_CLASSES SUPP_LHS
## 1 NCKAP5L,234817_at  control     value<cut,value<cut          1,1       18
## 2      MAP7,ATXN8OS  control     value>cut,value<cut          3,1       18
## 3      ZSCAN18,NPR2  control value<cut,cut<value<cut          1,2       19
##   SUPP_RHS ACC_RHS COV_LHS COV_RHS STAB_LHS STAB_RHS   CUT_1   CUT_2
## 1       18 0.97368 0.13740 0.30196        1        1 1.90584 1.64213
## 2       18 1.00000 0.13308 0.29932        1        1 2.51985 2.22742
## 3       19 0.98521 0.14616 0.32895        1        1 2.35647 2.54040
##     CUT_3 CUT_4         PVAL   RISK_PVAL REL_RISK    CONF_INT
## 1     NaN   NaN 4.818175e-06 0.005285147  2.28125 1.273:4.089
## 2     NaN   NaN 4.818175e-06 0.005285147  2.28125 1.273:4.089
## 3 2.59265   NaN 4.818175e-06 0.003949585  2.28125 1.298:4.009
```

Task 4: Exporting the rules to a text file

```
# saving file
saveLineByLine(sig_rules, "outputFile.txt")
```

Task 5: Using the VisuNet tool to upload the rules

As instructed, using the VisuNet tool at http://bioinf.icm.uu.se/~visunet/, the following image has been generated.



Task 6: Investigating the connections present on the networks

In the picture we can see that two clusters obatined one at the top and one at the bottom. Red nodes represent the genes that have high expression values and green that have low expression values. The nodes with thick blue outline represent that they have a large number of rules.

We observed that node PSMG4=3 is highlighted with most of red circles in bottom cluster and inter linked with other genes (e.g. SCIN=2, NCS1=2, NCS1=3). The MAP7=3 seems to be the most significant node in the top cluster. COX2=3 has the most number of rules. Most of the strongest connections appear in the bottom cluster.