Lab 5 - Gr. 14 - Bioinformatics (732A93)

Julius Kittler (julki092), Stefano Toffol (steto820), Saewon Jun (saeju204), Maximilian Pfundstein (maxpf364)

Task 1

Task:

- Install Packages

Task 2

Task: The sample data is stored in the workspace as autcon.

- Inspect the dataset and describe it.
- What is the number of features? What is the number of objects in each class?

Description:

- The number of features is 35 (excluding `decision`, the target variable, which is not a feature). All of them are numeric and represent genes.
- The number of rows is 146. Each row represents one male children with autism and healthy ones. There are 82 autistic boys and 64 control observations. This seems like a good enough balance of observations from both classes.

```
## Number of features:   35
```

```
## Number of observations:   146
```

```
## Number of obs. by class:
```

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

Task 3

Task: Run `rosetta()` on the default parameters: `autconDefault = rosetta(autcon)`

Use `autconDefault$main` to retrieve the rule table information, assign the result to a separate table. Use `autconDefault$quality` and display the quality statistics of the model:

a. Define what is cross-validation. How many cross-validations are performed in `rosetta` by default?
b. What is the default reduction method? What is it used for?
c. What is the default method of discretization? Describe it shortly. How many discretization bins are calculated?
d. What is the accuracy of the model?
e. How many rules do you obtain? Print the top three most significant rules. Which class get more significant rules? You can assume the rule to be significant if the p-value (PVAL) is lower than 0.05.

a)

b)

The default reduction method is Johnson. Different reducer methods are: Johnson, Genetic, Holte1R or Manual.

This describes the preparation, not the actual method. > The reduction method is used for finding the minimal subsets which preserves indiscernibility > between the samples. It is basically finding dependencies in the data. > So we're not looking for every combination of features but only at the combinations > that we actually find in the data.

c)

The default is `EqualFrequency`. It is used to transform continuous functions or data into discrete parts, also called bins. `EqualFrequency` divides the data into groups where each group has roughly the same size in terms of data-points/samples per bin. A method for finding the best number of bins is creating a histogram and using this the guess a good number.

For more information http://www.uta.fi/sis/tie/tl/index/Datamining6.pdf.

d)

The accuracy can be taken from the confusion matrix which shows the four classifications: - control as control (TP) - control as autism (FP) - autism as control (FN) - autism as autism (TN)

The accuracy is calculated by: (TP+TN)/(TP+TN+FP+FN).

It basically tells us the misclassification rate of our model and the α and β errors made.

e)

```
## We got 191 rules.
```

FEATURES	DECISION	CUTS_COND	DISC_CLASSES	SUPP_LHS	SUPP_RHS	ACC_
NCKAP5L,234817_at	control	value<cut,value<cut	1,1	18	18	0.9
MAP7,ATXN8OS	control	value>cut,value<cut	3,1	18	18	1.0
ZSCAN18,NPR2	control	value<cut,cut<value<cut	1,2	19	19	0.9

```
##
##  autism control
##     108      77
```

Task 4

Task: Export the rules to a text file using the `saveLineByLine()` function.

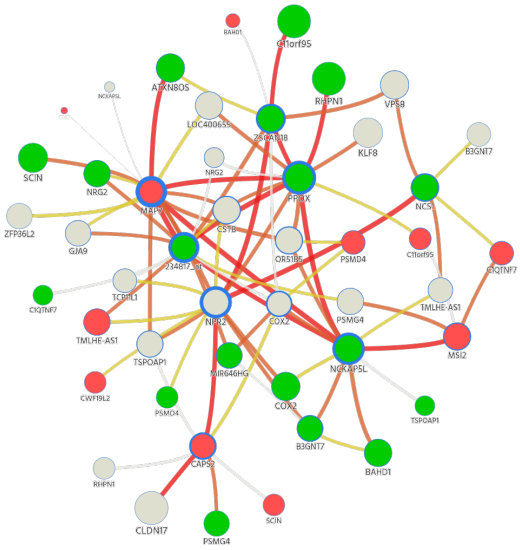

Figure 1: VisuNet - Control

Task 5

Task: Choose further options: Use the VisuNet tool at http://bioinf.icm.uu.se/~visunet/. Upload your rules.

> File format Choose: "Line by line" Minimum Accuracy Default is 0.7, which means that rules with at least 70% accuracy will be used for displaying a network. Minimum Support Default is 1, which means all rules will be included in the network. You can toggle that and see the effects on the network. Threshold (%) Keep it 100 Show top n nodes Leave it blank Color of nodes Choose: "Level of the gene expression" Is this gene data? Yes. Use the autism_annot.txt file

Submit the file to generate a rule-based network. On the left side, there is "Information Bar" where you can select the decision. By clicking on a node, you will be able to see:

a. In the bottom panel: the rules that have the node as one of its conditions.
b. On the right-hand side in the "Selected Node/Edge Info": the name of the node, the number of edges it has, the mean accuracy value and the mean support value. Furthermore, there is information about KEGG Pathways and GO annotations related with the node.

Task 6

Export the networks for the autism and control decisions. Investigate connections present on the networks. Find the strongest connections and the most significant nodes for each decision. Try to interpret these in the context of autism related genes. Hints: Calcium homeostasis is altered in autism disorders (Palmieri et al., 2010). The autism dataset includes a group of genes related to a calcium ion binding (GO:0005509). Take a closer look at SCIN, NCS1 and CAPS2. You may also use the SFARI GENE database containing information about autism-related genes: https://gene.sfari.org/

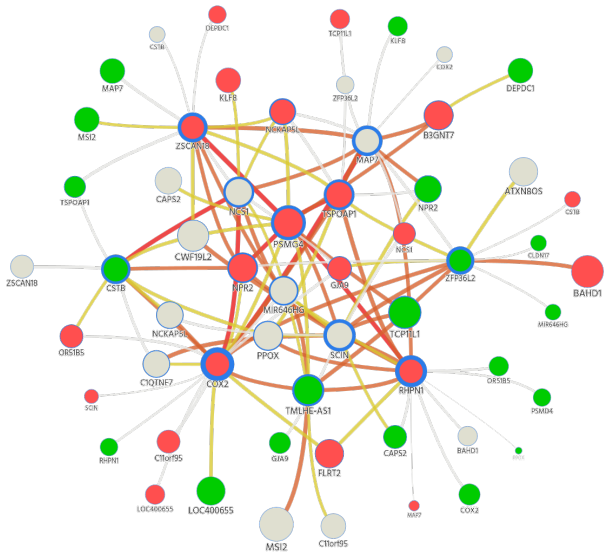

Figure 2: VisuNet - Autism

Appendix

```
knitr::opts_chunk$set(fig.width = 7, fig.height = 3, echo = FALSE,
                      warning = FALSE, message = FALSE)

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


cat("Number of features:  ", ncol(autcon)-1, "\n")
cat("Number of observations:  ", nrow(autcon), "\n")
cat("Number of obs. by class:  \n")
table(autcon$decision)

# -----------------------------------------------------------------------------
# Task 3
# -----------------------------------------------------------------------------

autconDefault = rosetta(autcon)
rule_table = autconDefault$main
quality_statistics = autconDefault$quality



cat("We got", nrow(rule_table), "rules.")
knitr::kable(head(rule_table[order(rule_table$PVAL),], 3),
             caption = "Top 3 Significant Rules")
```

```r
#kable(autconDefault$rules[1:3])

rule_table_sub = rule_table[which(rule_table$PVAL < 0.05), ]
table(rule_table_sub$DECISION)


saveLineByLine(autconDefault$main, "outFile.txt")
```