Rule-based classification and visualization

Instructors:
Mateusz Garbulowski mateusz.garbulowski@icm.uu.se
Karolina Smolinska karolina.smolinska@icm.uu.se

Background

R.ROSETTA is an R package implemented as a wrapper around the ROSETTA rough set-based system. In addition to all the existing ROSETTA algorithms, we have added new functions especially useful in bioinformatics applications. These include: undersampling, rule p-value estimation, rule visualization methods and detection of support sets. The package and installation manual are publicly available on the GitHub:
https://github.com/mategarb/R.ROSETTA.

R.ROSETTA includes 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. The following decision table contains features selected by a fast correlation-based filter (FCBF). The example data is publicly available at GEO repository with the reference number GSE25507. More information about the data can be found on the NCBI webpage: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25507.

Tasks

1. Open RStudio and install R.ROSETTA from GitHub repository. Load the package.

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

2. 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? Do you think the distribution of objects is balanced?

   Hints:
   - you can use the `View()` function to display the data:
     `View(autcon)`
   - Function `table()` can be useful to sum up the decision.

3. 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.

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

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

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.

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/