R.ROSETTA: a package for analysis of rule-based classification models

Mateusz Garbulowski[1,*], Klev Diamanti[1,#], Karolina Smolińska[1,#], Patricia Stoll[2], Susanne Bornelöv[3], Aleksander Øhrn[4] and Jan Komorowski[1,*]

[1]Department of Cell and Molecular Biology, Uppsala University, Sweden

[2]Department of Biosystems Science and Engineering, ETH Zurich, Switzerland

[3]Wellcome Trust Medical Research Council Stem Cell Institute, University of Cambridge, England

[4]Department of Informatics, Oslo University, Norway

[#]These authors contributed equally to the work as second authors.

*Corresponding authors

E-mail: mateusz.garbulowski@icm.uu.se (Mateusz Garbulowski)

E-mail: jan.komorowski@icm.uu.se (Jan Komorowski)

R.ROSETTA is freely available at: https://github.com/komorowskilab/R.ROSETTA

Abstract

ROSETTA is a rough set-based classification toolkit that aims at identifying semantics from various data types. Here we present the R.ROSETTA package, which is an R wrapper of ROSETTA. The package significantly enhances the accessibility of the existing machine learning environment and the interpretability of the results. The ROSETTA functions have been enriched and improved by the incorporation of novel components targeting bioinformatics applications. Such improvements include: undersampling imbalanced datasets, estimation of the statistical significance of classification rules, retrieval of support sets, prediction of external data and integration with rule visualization frameworks. We tested the performance of R.ROSETTA on a complex dataset involving gene expression measurements for autistic and non-autistic young males. We demonstrated that R.ROSETTA facilitated the detection of novel gene-gene interactions. The results demonstrated the potential of R.ROSETTA classifiers to identify putative biomarkers and novel biological interactions.

Keywords: rough sets, machine learning, rule-based model, statistical significance of models and rules

Introduction

Rough set classification is a transparent machine learning technique that has been widely applied in various scientific areas (Kumar and Inbarani, 2018; Zhang et al., 2014). The rough set methodology creates rule-based classification models that consist of minimal sets of IF-THEN rules that uncover interactions among variables (Pawlak, 1982). The ROSETTA software is an implementation of rule-based classification modeling (Øhrn and Komorowski, 1997). The framework has been used to solve biology-related issues e.g. (Gil-Herrera et al., 2011; Komorowski, 2014; Setiawan et al., 2009). Here we present a more accessible and flexible implementation of ROSETTA in a form of an R package. R.ROSETTA substantially extends the functionality of the existing software towards analyzing complex and ill-defined bioinformatics datasets. Among others, we have implemented functions (Figure 1) such as undersampling, rule p-value estimation, class prediction, support sets retrieval and rule visualization approaches. To evaluate R.ROSETTA performance, we explored rule-based models for a complex dataset of gene expressions measured for autistic and non-autistic samples (Supplementary Table S1).

Implementation

R.ROSETTA was implemented under R (R Core Team, 2018) version 3.5.2 and the open-source R package is available on GitHub repository at https://github.com/mategarb/R.ROSETTA. The R.ROSETTA package is a wrapper (Supplementary Note – Package architecture) around command line version of ROSETTA system. In contrast to ROSETTA, R.ROSETTA is a cross-platform application with additional functionalities (Figure 1). In the supplementary material, we are providing an overview of the main upgrades (Supplementary Note – Main upgrades) introduced in R.ROSETTA.

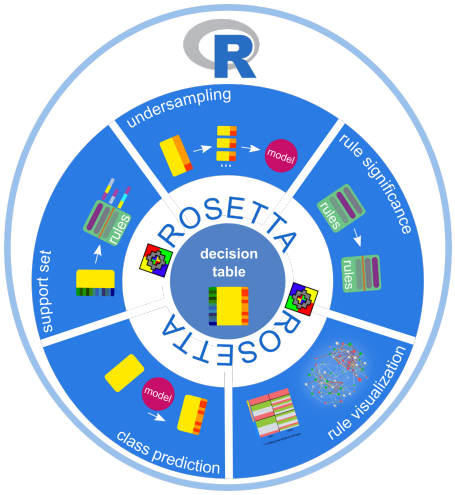

Fig. 1. Overview of R.ROSETTA and the major components that were implemented to enhance the ROSETTA functionality.

Results

We examined gene expression levels (Alter et al., 2011) of autistic and non-autistic (control) male children (Supplementary Table S1). The dataset has been preprocessed (Supplementary Note – Data preprocessing) and corrected for subject age effect (Supplementary Figure S1). In the next step, we employed the Fast Correlation-Based Filter dimensionality reduction method (Novoselova et al., 2018) (Supplementary Note - Feature selection, Supplementary Table S2). The decision table was reduced to 35 features (Supplementary Table S3) out of which 11 have been previously linked to autism disorder (Supplementary Note – Feature validation; Supplementary Figure S2).

We constructed models (Supplementary Note – Classification) in R.ROSETTA for Johnson and Genetic reducers with 82% and 91% accuracy, respectively (Supplementary Table S4, Supplementary Table S5). Even though the overall performance of the Genetic algorithm was better than Johnson's, its tendency of generating numerous rules reduced the significance of individual rules after correcting for multiple testing (Supplementary Figure S3; Supplementary Table S4).

We selected highly significant rules from the Johnson model to identify the most important genes for each decision class separately. Among the 15 genes occurring in the rules 10 genes are likely to be associated to autism (Supplementary Table S6) e.g. elevated expression of COX2 has been earlier identified with autism (Yoo et al., 2008). TSPOAP1 has been associated (Bucan et al., 2009) with autism through a deletion in an exonic locus. Furthermore, we identified genes related to calcium homeostasis control such as NCS1 and SCIN. Previous studies (Palmieri et al., 2010) have demonstrated that calcium homeostasis is altered in autism disorders. We detected that expression of antisense RNA of TMLHE (TMLHE-AS1) is down-regulated (Supplementary Figure S2). The TMLHE gene is a well-known risk factor of autism (Celestino-Soper et al., 2011). We discovered also a zinc finger gene ZFP36L2. The association of zinc fingers to autism was previously described by the dataset authors (Alter et al., 2011).

Finally, two strongest interactions (Supplementary Table S7) for autism-related rules contained genes previously linked to autism or its symptoms. Our study showed (Supplementary Figure S4) that up-regulated expression of TSPOAP1 was associated with up-regulated expression of PSMG4. The second interaction in autism class consist of unchanged expression of NCS1 and down-regulated expression of CSTB. The reduced expression of CSTB has been linked to the mechanism of pathogenesis in epilepsy (Lalioti et al., 1997).

Conclusions

R.ROSETTA is a tool that gathers fundamental components of statistics for rule-based modelling. Additionally, the package provides hypotheses about potential interactions between features that discern phenotypic classes.

Acknowledgment

We would like to thank N. Baltzer, F. Barrenäs, M. Cavalli, Z. Khaliq, B. T. Moghadam, G. Pan, C. Wadelius and S. Younes for their insightful discussions and testing/debugging the package.

Conflict of Interest: none declared.

References

Alter, M.D. *et al.* (2011) Autism and increased paternal age related changes in global levels of gene expression regulation. *PloS one*, **6.2**, e16715.

Bucan, M. *et al.* (2009) Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes. **5.6**, e1000536.

Celestino-Soper, P.B. *et al.* (2011) Use of array CGH to detect exonic copy number variants throughout the genome in autism families detects a novel deletion in TMLHE. **20.22**, 4360-4370.

Gil-Herrera, E. *et al.* (2011) Rough set theory based prognostication of life expectancy for terminally ill patients. Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE, 6438-6441.

Komorowski, J. (2014) Learning rule-based models-the rough set approach. *Amsterdam: Comprehensive Biomedical Physics*.

Kumar, S.S. and Inbarani, H.H. (2018) Cardiac arrhythmia classification using multi-granulation rough set approaches. *International Journal of Machine Learning Cybernetics*, **9.4**, 651-666.

Lalioti, M.D. *et al.* (1997) Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. *Nature*, **386.6627**, 847.

Novoselova, N. *et al.* (2018) Biocomb: Feature Selection and Classification with the Embedded Validation Procedures for Biomedical Data Analysis. Release R Package Version 0.4. https://CRAN.R-project.org/package=Biocomb. (1 October 2018 last accessed).

Øhrn, A. and Komorowski, J. (1997) ROSETTA--A Rough Set Toolkit for Analysis of Data. Proc. Third International Joint Conference on Information Sciences.

Palmieri, L. *et al.* (2010) Altered calcium homeostasis in autism-spectrum disorders: evidence from biochemical and genetic studies of the mitochondrial aspartate/glutamate carrier AGC1. *Molecular psychiatry*, **15.1**, 38.

Pawlak, Z. (1982) Rough sets. *International journal of computer information sciences*, **11.5**, 341-356.

R Core Team (2018) R: A language and environment for statistical computing., R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Setiawan, N.A. *et al.* (2009) Diagnosis of coronary artery disease using artificial intelligence based decision support system. proceedings of the international conference on man-machine systems (ICoMMS), Batu Ferringhi, Penang.

Yoo, H.J. *et al.* (2008) Association between PTGS2 polymorphism and autism spectrum disorders in Korean trios. *Neuroscience research*, **62.1**, 66-69.

Zhang, J. *et al.* (2014) A comparison of parallel large-scale knowledge acquisition using rough set theory on different MapReduce runtime systems. *International Journal of Approximate Reasoning*, **55.3**, 896-907.