

GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest

Ramón Diaz-Uriarte*¹

¹Statistical Computing Team, Structural Biology and Biocomputing Programme, Spanish National Cancer Center (CNIO), Melchor Fernández Almagro 3, Madrid, 28029, Spain

Email: Ramon Diaz-Uriarte* - rdiaz02@gmail.com;

*Corresponding author

Abstract

Background: Microarray data are often used for patient classification and gene selection. An appropriate tool for end users and biomedical researchers should combine user friendliness with statistical rigor, including carefully avoiding selection biases and allowing analysis of multiple solutions, together with access to additional functional information of selected genes. Methodologically, such a tool would be of greater use if it incorporates state-of-the-art computational approaches and makes source code available.

Results: We have developed GeneSrF, a web-based tool, and varSelRF, an R package, that implement, in the context of patient classification, a validated method for selecting very small sets of genes while preserving classification accuracy. Computation is parallelized, allowing to take advantage of multicore CPUs and clusters of workstations. Output includes bootstrapped estimates of prediction error rate, and assessments of the stability of the solutions. Clickable tables link to additional information for each gene (GO terms, PubMed citations, KEGG pathways), and output can be sent to PaLS for examination of PubMed references, GO terms, KEGG and Reactome pathways characteristic of sets of genes selected for class prediction. The full source code is available, allowing to extend the software. The web-based application is available from <http://genesrf.bioinfo.cnio.es>. All source code is available from Bioinformatics.org (<http://bioinformatics.org/asterias>) or The Launchpad (<http://launchpad.net/asterias>). The R package is also available from CRAN (<http://cran.r-project.org>).

Conclusions: varSelRF and GeneSrF implement a validated method for gene selection including bootstrap estimates of classification error rate. They are valuable tools for applied biomedical researchers, specially for exploratory work with microarray data. Because of the underlying technology used (combination of parallelization with web-based application) they are also of methodological interest to bioinformaticians and biostatisticians.

Background

Patient classification and gene selection related to classification are common uses of microarray data (e.g., review and references in [1]), but statistically rigorous and user-friendly tools for gene selection in the context of class prediction are rare. Such a tool should address two major issues. First, it should provide unbiased estimates of the prediction error rate of the procedure. Most users are by now aware of “selection bias”, as originally reported in [2,3], but bias caused by trying different methods and/or sets of genes, and choosing the one with the smallest cross-validated error rate [4] is still not widely recognized. In this later case we need a nested [4] or double or full cross-validation [5] to estimate the error rate of the rule or procedure. Second, we need to assess the so called multiplicity (or lack of uniqueness) problem: variable selection with microarray data can lead to many solutions that have similar prediction errors, but that share few common genes [1,6–9]. Choosing any one particular set of genes without being aware of the variability in solutions can lead to a false sense of certainty in the selected set.

From a users’ perspective, an ideal tool should also be user friendly and provide additional resources to ease the interpretation of results. Web-based tools are an excellent platform as they do not require software installation or upgrades from the user. In addition, web based tools, if appropriately implemented, allow easy access to information such as Gene Ontology terms, the UCSC and Ensembl databases, KEGG and Reactome pathways, or PubMed references, thus enhancing the biological interpretation of results. Moreover, web-based tools, if implemented appropriately, can harvest computational resources rarely available to most individual users, specifically the increasing availability of multicore processors and easily accessible clusters made with off-the-shelf components. Currently, the major opportunities for improved performance as well as the ability to analyze ever larger data sets do not lie in faster CPUs but in being able to use parallel and distributed computing to exploit multi-core servers and clusters [10,11]. In

addition to providing a benefit to the end user (decreased execution time), tools that combine parallelization with web-based programming are important methodological developments. Finally, a tool that fulfills the above requirements is of much greater relevance if it makes its source code available under an open-source license. Source code availability allows the research community to experiment with, and improve upon, the method and fix bugs, encourages reproducible research, allows to verify claims by method developers, makes the international research community the owner of the tools needed to carry out its work and, thus, creates the conditions for swift progress upon previous work, concerns of particular importance in bioinformatics [12,13].

We have developed GeneSrF and varSelRF (a web-based application and R package, respectively), that satisfy the above requirements. The only available web-based tools with similar scope are M@CBETH [14] and Prophet [15]. These tools, however, do not examine the multiplicity problem, cannot benefit from multicore processors or computing clusters, and do not make source code available. M@CBETH, in addition, is restricted to two-class problems and does not focus on the gene selection problem. Prophet, in turn, does not seem to solve satisfactorily the biased error rate problem (it reports the error rate as that of the classifier with smallest cross-validated error rate, without evaluating the error rate of the rule itself).

Implementation

The core statistical functionality is provided by the varSelRF package for R [16]. This package implements the procedure in [1] for gene selection using random forests, building upon the randomForest package <http://cran.r-project.org/src/contrib/Descriptions/randomForest.html>, an R port by A. Liaw and M. Wiener of the original code by L. Breiman and A. Cutler. We use MPI [17] for parallelization via the R-packages Rmpi (<http://www.stats.uwo.ca/faculty/yu/Rmpi>), by H. Yu, and Snow (<http://cran.r-project.org/src/contrib/Descriptions/snow.html>), by L. Tierney, A. J. Rossini, Na Li and H. Sevcikova. In the web-based application, the CGI, initial data validation, and the setting-up and closing of the parallel infrastructure (booting and halting the LAM/MPI universes) is implemented with Python. Our installation runs on a cluster of 30 dual processor workstations.

The web-based application provides tables with clickable links to our application IDClight (<http://idclight.bioinfo.cnio.es>) [18], that provides mappings between gene and protein identifiers, PubMed references, Gene Ontology terms, and KEGG and Reactome pathways, for the selected gene. The web-based application also allows the user to send the sets of selected genes to our tool PaLS (<http://idclight.bioinfo.cnio.es>), to examine PubMed references, Gene Ontology terms, KEGG pathways or

Reactome pathways that are common to a user-selected percentage of genes or lists (bootstrap solutions). The input for the web-based application are either plain text files, or files that come from other tools of the Asterias suite (<http://asterias.bioinfo.cnio.es>). GeneSrF has been running in production use for over a year. Further documentation and examples for the web-based application are available from its on-line help, and for the R package from the standard R documentation system. Bug-tracking and additional tests are available from Bioinformatics.org and The Launchpad.

Results and discussion

Our procedure is explicitly targeted to select very small sets of genes, and has been shown [1] to have a classification error rate on-par with other, state-of-the-art, classification procedures. Additionally, our programs allow the exploratory usage of random forest for identifying large subsets of genes potentially relevant for class prediction. In contrast to other tools, such as M@CBETH [14], we are not restricted to two-class problems.

To avoid underestimating the error rate of the classification procedure, we use the bootstrap (the 0.632+ approach of [19]). As in [1], we bootstrap the complete procedure, including selecting the classifier with minimal out-of-bag error rate (thus, this is a “full” or “double” bootstrap procedure, sensu [5]), and thus our estimates of error rate are not affected by selection biases. This contrasts, for instance, with Prophet [15], where the error rate reported is that the classifier with smallest cross-validated error rate. Based upon the bootstrap results, we also show the average out-of-bag predictions for each sample, allowing to easily assess poorly predicted samples and potential outliers. There are other tools available for performing cross-validation and bootstrap of classification methods, such as the R package `ipred` (<http://cran.r-project.org/src/contrib/Descriptions/ipred.html>) by A. Peters and T. Hothorn, or the BioConductor package `MCRestimate` (<http://bioconductor.org/packages/1.9/bioc/html/MCRestimate.html>), by M. Ruschhaupt, U. Mansmann, P. Warnat, W. Huber and A. Benner, specifically targeted to computing misclassification error rates combining the gene selection and classification steps. Our approach, however, has been tailored to our own variable selection procedure and has been parallelized.

A unique feature of GeneSrF and varSelRF are their emphasis on examining possible multiple solutions. Since we obtain 200 resamples in the process of bootstrapping (see above) there is little added computational cost to providing analysis of stability and multiplicity of solutions. We report the number of genes selected and the identity of the individual genes selected in the original sample and the 200

bootstrap runs, including frequencies of every gene selected in the solutions. Moreover, the biological interpretation of the results is enhanced by the access to additional information. For each gene in the results, a link to IDClight (<http://idclight.bioinfo.cnio.es>) [18], allows the user to obtain additional information, including mapping between gene and protein identifiers, PubMed references, Gene Ontology terms, and KEGG and Reactome pathways. The multiple solutions can be further studied by sending sets of selected genes to our tool PaLS (<http://idclight.bioinfo.cnio.es>) to examine PubMed references, Gene Ontology terms, KEGG pathways, or Reactome pathways that are common to a user-selected percentage of genes or lists (bootstrap solutions). Some sample output is shown in Figure 1.

Finally, GeneSrF is one of the very few tools for the analysis of gene expression data that uses parallelization and, as far as we know, the only web-based tool to use parallelization for gene selection and classification. This is an important methodological novelty, as we can no longer expect that increases in single-CPU speed will allow us to analyze larger data sets in shorter time: the rate of increase in CPU speed has slowed down considerably in the last five years but, in contrast, increasing numbers of CPU cores (either in individual machines —including laptops— or via off-the-shelf computing clusters) are becoming much more affordable [10, 11]. Thus, further decreases in user wall time (time to wait for a result) and ability to tackle more complex problems will depend on our ability to use parallel, distributed, and concurrent programming. GeneSrF therefore represents a case example on combining parallel computing with a user-friendly web-based application for the analysis of gene expression data and, by making the full source code available, allows other researchers to build upon our developments.

Future work focuses on extending the software to use random forest-related techniques applicable to heterogeneous types of variables such as addition of categorical data [20] and other clinical information. As well, we are exploring alternative mechanisms and languages for parallelizing and distributing computations, and we are rewriting most of the code using Pylons (<http://pylonshq.com>), a Python web framework, to try to simplify installation of the web-based application.

Conclusion

varSelRF and GeneSrF implement a validated method for gene selection and provide bootstrap estimates of classification error rate, take advantage of computing clusters and multicore processors, and encourage careful examination of the multiplicity of solutions problems. Thus, these are both useful tools for applied biomedical researchers using microarray and gene expression data, and represent unique methodological developments in the area of web-based gene expression analysis tools.

Availability and requirements

For GeneSrF:

Project name: GeneSrF

project home page: <http://genesrf.bioinfo.cnio.es>

Operating system: Platform independent (web-based application)

Programming language: R, Python

Other requirements: A web browser.

License: None for usage. Web-based code: Affero GPL (open source).

Any restrictions to use by non-academics: None.

For varSelRF:

Project name: varSelRF

project home page: <http://launchpad.net/varself>

Operating system: Linux, Unix

Programming language: R, Python

Other requirements: LAM/MPI

License: GNU GPL

Any restrictions to use by non-academics: None

List of abbreviations used

CGI Common Gateway Interface

GO Gene Ontology

KEGG Kyoto Encyclopedia of Genes and Genomes

LAM Local Area Multicomputer

MPI Message Passing Interface

Acknowledgements

A. Alibés and A. Cañada for their work on IDClight and PaLS. Bioinformatics.org and The Launchpad for project and repository hosting. Funding provided by Fundación de Investigación Médica Mutua Madrileña and Project TIC2003-09331-C02-02 of the Spanish Ministry of Education and Science (MEC). R.D.-U. is partially supported by the Ramón y Cajal programme of the Spanish MEC.

References

1. Díaz-Uriarte R, Alvarez de Andrés S: **Gene selection and classification of microarray data using random forest.** *BMC Bioinformatics* 2006, **7**.
2. Ambroise C, McLachlan GJ: **Selection bias in gene extraction on the basis of microarray gene-expression data.** *Proc Natl Acad Sci USA* 2002, **99**(10):6562–6566.
3. Simon R, Radmacher MD, Dobbin K, McShane LM: **Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification.** *Journal of the National Cancer Institute* 2003, **95**:14–18.
4. Varma S, Simon R: **Bias in error estimation when using cross-validation for model selection.** *BMC Bioinformatics* 2006, **7**.
5. Dudoit S, Fridlyand J: **Classification in microarray experiments.** In *Statistical analysis of gene expression microarray data*. Edited by Speed T, New York: Chapman & Hall 2003:93–158.
6. Somorjai RL, Dolenko B, Baumgartner R: **Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions.** *Bioinformatics* 2003, **19**:1484–1491.
7. Pan KH, Lih CJ, Cohen SN: **Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays.** *Proc Natl Acad Sci U S A* 2005, **102**:8961–8965.
8. Ein-Dor L, Kela I, Getz G, Givol D, Domany E: **Outcome signature genes in breast cancer: is there a unique set?** *Bioinformatics* 2005, **21**:171–178.
9. Michiels S, Koscielny S, Hill C: **Prediction of cancer outcome with microarrays: a multiple random validation strategy.** *Lancet* 2005, **365**:488–492.
10. Sutter H: **The Free Lunch Is Over: A Fundamental Turn Toward Concurrency in Software.** *Dr. Dobbs's Journal* 2005, **30**(3):202–210.
11. Kontoghiorghes EJ: *Handbook of Parallel Computing and Statistics*. Boca Raton, FL: Chapman & Hall, CRC 2006.
12. Dudoit S, Gentleman RC, Quackenbush J: **Open source software for the analysis of microarray data.** *Biotechniques* 2003, **Suppl**:45–51.
13. Díaz-Uriarte R: **Supervised methods with genomic data: a review and cautionary view.** In *Data analysis and visualization in genomics and proteomics*. Edited by Azuaje F, Dopazo J, New York: Wiley 2005:193–214.
14. Pochet NL, Janssens FA, De Smet F, Marchal K, Suykens JA, De Moor BL: **M@CBETH: a microarray classification benchmarking tool.** *Bioinformatics* 2005, **21**:3185–3186.
15. Medina I, Montaner D, Tarraga J, Dopazo J: **Prophet, a web-based tool for class prediction using microarray data.** *Bioinformatics* 2006, **in press**.
16. R Development Core Team: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria 2004, [<http://www.R-project.org>]. [ISBN 3-900051-00-3].
17. Pacheco P: *Parallel programming with MPI*. San Francisco: Morgan kaufman 1997.
18. Alibés A, Yankilevich P, Cañada A, Diaz-Uriarte R: **IDconverter and IDClight: conversion and annotation of gene and protein IDs.** *BMC Bioinformatics* 2007, **8**.
19. Efron B, Tibshirani RJ: **Improvements on cross-validation: the .632+ bootstrap method.** *J. American Statistical Association* 1997, **92**:548–560.
20. Strobl C, Boulesteix AL, Zeileis A, Hothorn T: **Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution.** *BMC Bioinformatics* 2007, **8**:25.

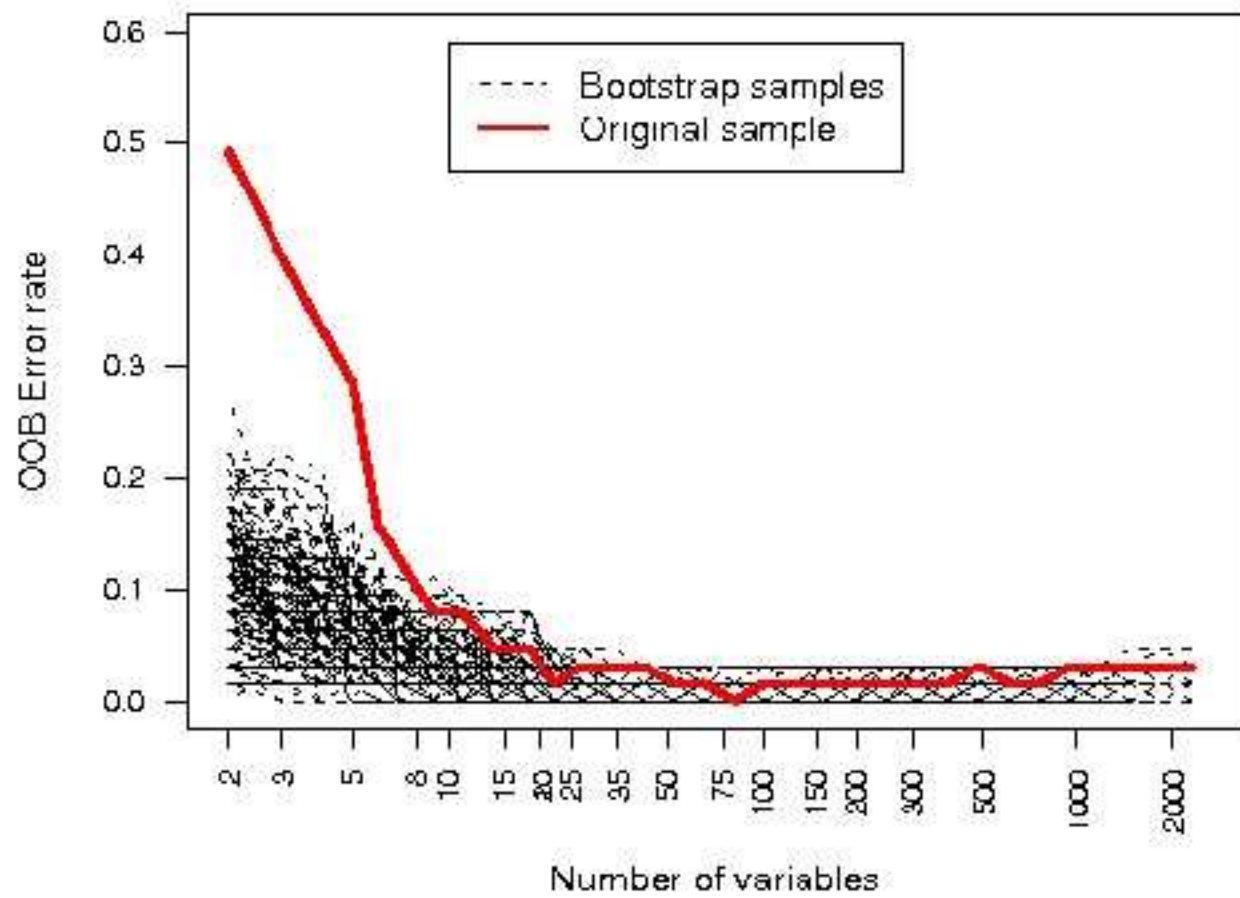
Figures

Figure 1 - Example output

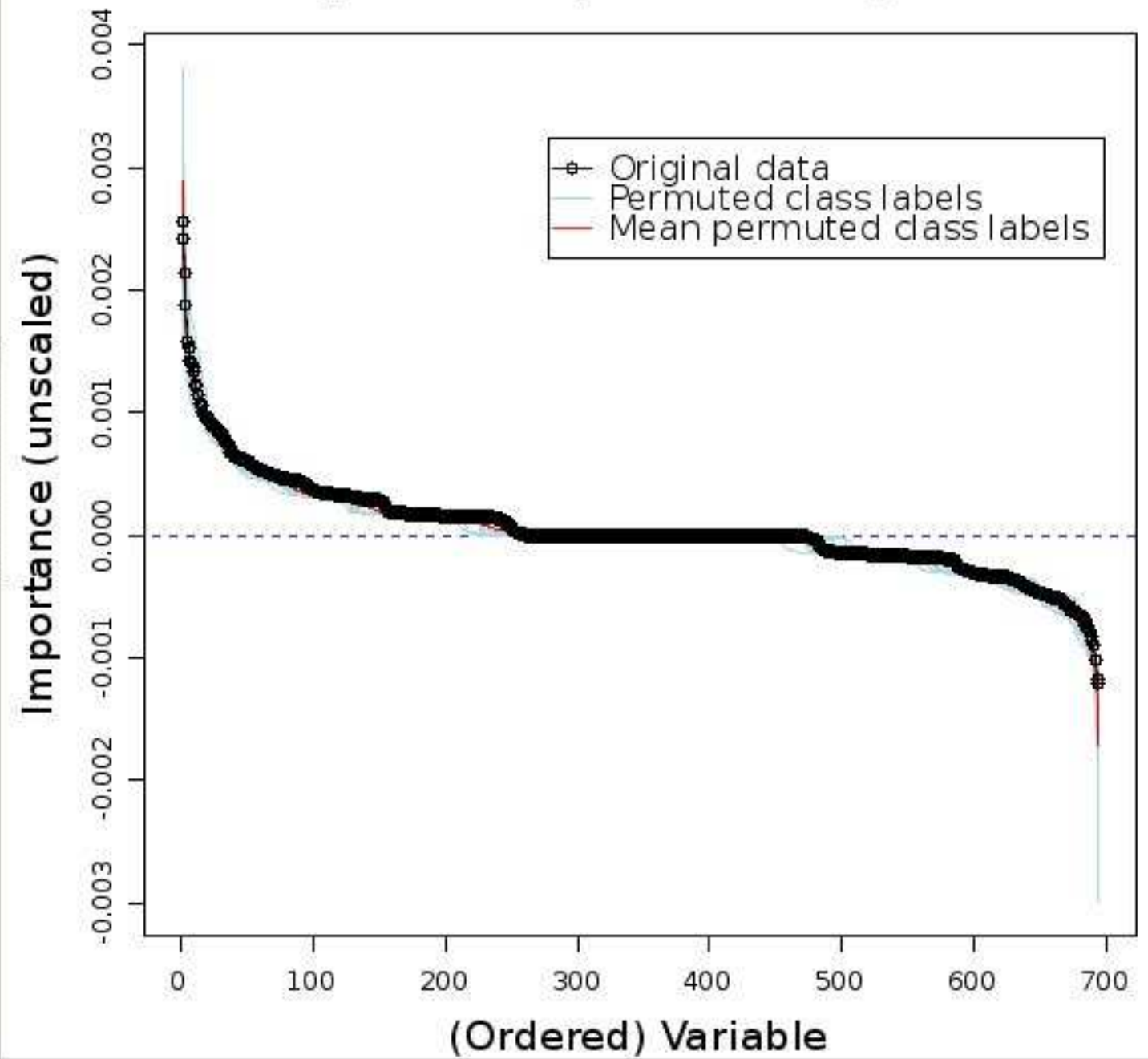
Background and left: relationship between error rate and number of genes used for prediction and plots of out-of-bag predictions, in a four-class example with relatively good predictive accuracy. Upper-right: variable importance spectrum in a model where no prediction is possible.

OOB error vs. n

OOB Error rate vs. Number of variables in predictor

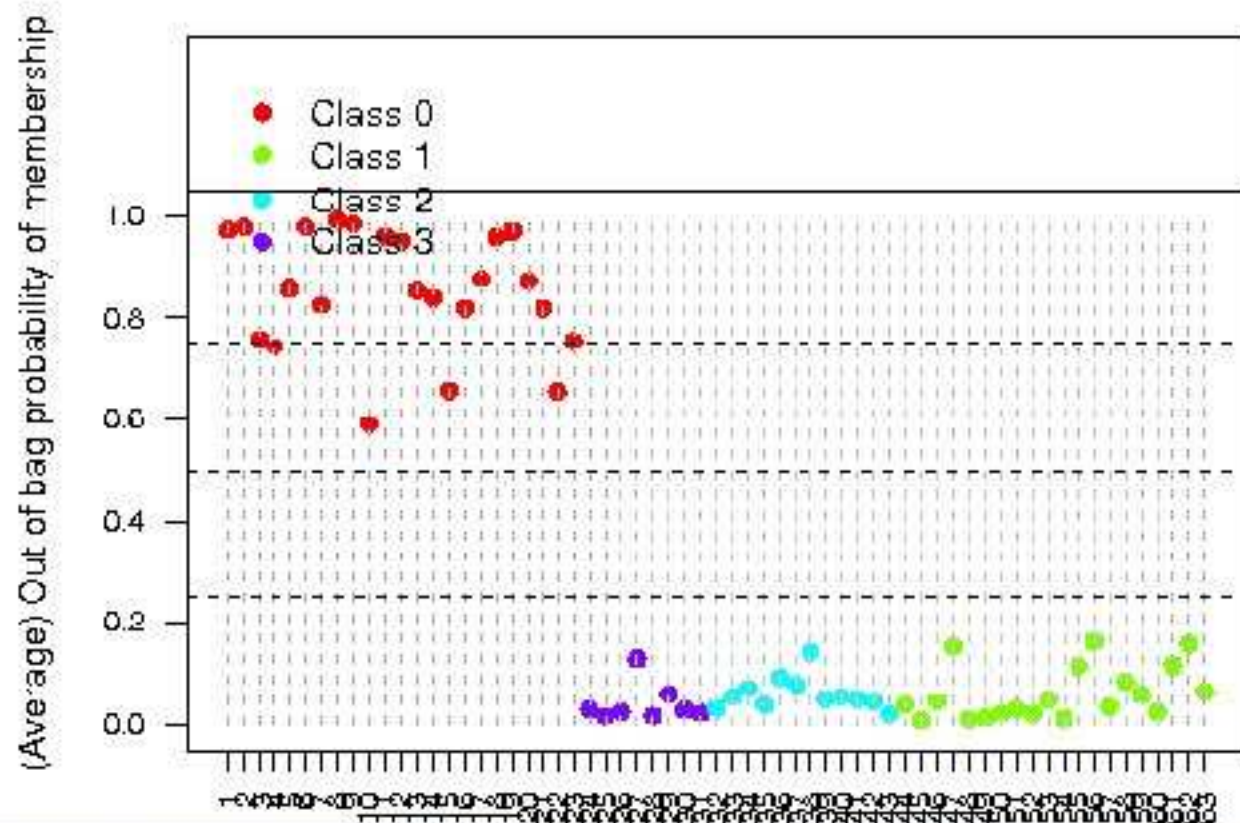


Importance Spectrum: all genes



OOB predictions [\(help\)](#)

Class 0



Class 1

