Assessment of the performance of advanced classification models applied for concentration of trace elements in healthy and cancerous prostate tissue

K. Banas[1], A. Banas[1], B. Pawlicki[2], A. Pawlicka[3] , M. Gajda[4], G. Dyduch[5], W. M. Kwiatek[6], and M. B. H. Breese[1,7]

[1]Singapore Synchrotron Light Source, National University of Singapore, 5 Research Link, Singapore 117603
[2]Gabriel Narutowicz Hospital, Pradnicka 37, 31-202 Krakow, Poland
[3]School of Medicine in English, Jagiellonian University Medical College, sw. Anny 12, 31-008 Krakow, Poland
[4]Department of Histology, Collegium Medicum, Jagiellonian University, Kopernika 9, 31-034 Krakow, Poland
[5]Department of Pathomorphology, Collegium Medicum, Jagiellonian University, Grzegorzecka 16, 31-034 Krakow, Poland
[6]Institute of Nuclear Physics PAN, Radzikowskiego 152, 31-342 Krakow, Poland
[7]Physics Department, National University of Singapore, 2 Science Drive 3, Singapore 117542

INTRODUCTION

Gleason grade is the most popular system used to classify different stages of prostate cancer. However, this method is highly dependent on experience of the histopathologists as the assessment is based only on microscopic appearance of analysed tissue sections. Complimentary system is needed that integrates tissue architecture and its biochemistry. SRIXE (Synchrotron Radiation Induced X-ray Emission) is powerful, yet not destructive technique used to determine the elemental make-up of a sample. As very often valuable information is hidden in a huge number of variables, that is why dedicated multivariate statistical method has to be applied for straightforward analysis. Comprehensive comparison of classification models based on elemental concentrations in prostate tissue is presented in this poster.

SAMPLES, DATA and METHODS

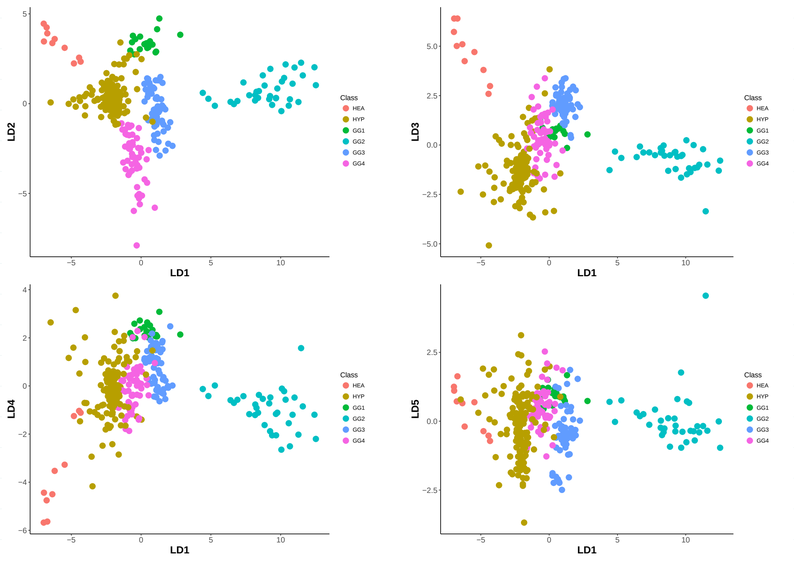
Samples were prepared in a form of thin slices (14 μm) placed on mylar foil; adjacent sections for histopathological analysis were put on microscopic glass. Tissue sections were taken from 12 patients who underwent radical prostatectomy. Spectra were collected at the beamline L (Hasylab, DESY). White beam of 17 keV was used to irradiate the samples. Data sets were analysed by using R statistical platform, open source code-driven and scalable solution within RStudio graphic user interface with additional packages loaded: dplyr (for data manipulation), FactoMineR (for dimension reduction and clustering), ggplot2 (for visualization) and mixOmics (for PLS-DA testing and tuning).

CLUSTERING TECHNIQUES

Various multivariate statistical methods are tested for dimension reduction, clustering and identification. Both unsupervised (principal component analysis (PCA), hierarchical cluster analysis (HCA), partitioning around medoids (PAM) and k-means) and supervised (linear discriminant analysis (LDA), partial least squares discriminant analysis (PLS-DA) techniques are evaluated in terms of their performance benchmark and computational time.

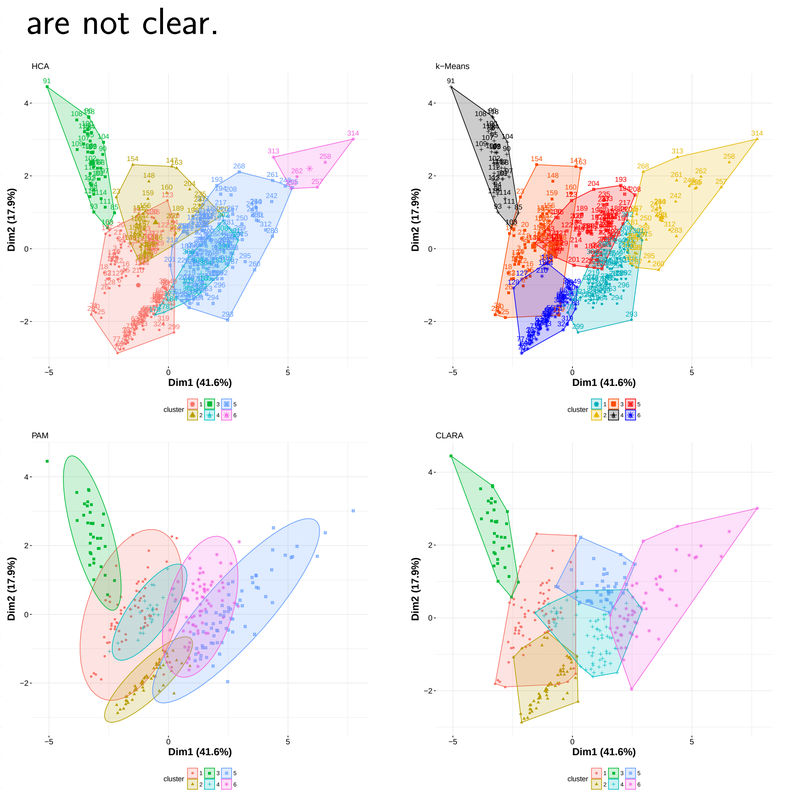
LDA

Linear discriminant analysis (LDA) is classical method for supervised learning. By using the linear combinations of original variables (elemental concentrations) classification model may be constructed.



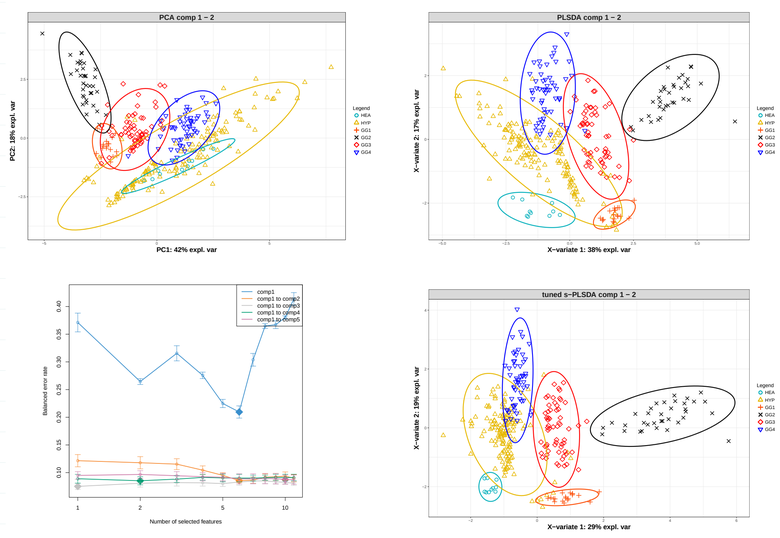
HCA, k-means, PAM, CLARA

Outputs for various unsupervised clustering methods: k-means, HCA, PAM and CLARA are presented in the plots for the first two components. In every case some cluster structure can be observed in the data set, but the boundaries between the clusters are not clear.



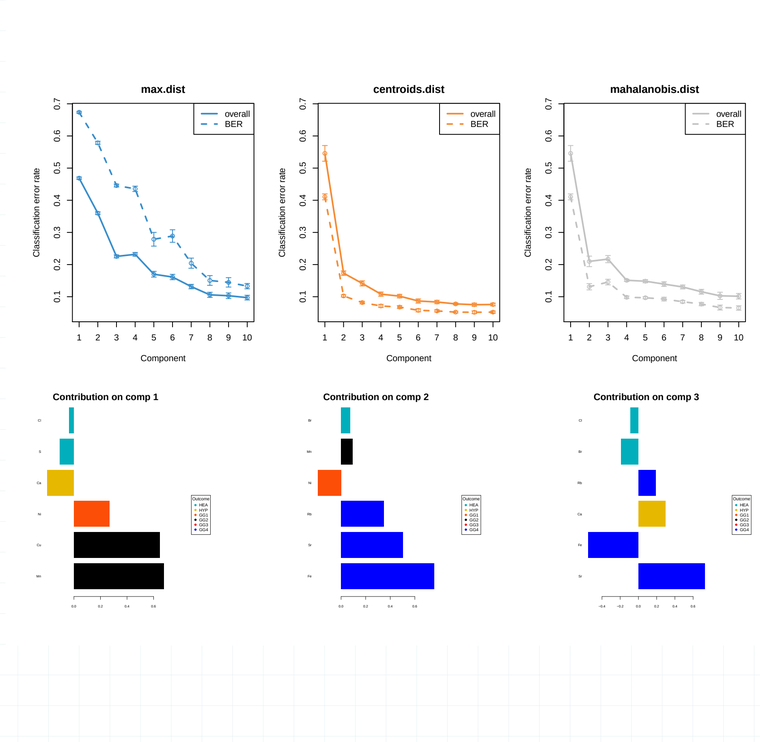
PLS-DA

PLS-DA is an example of supervised analysis. Comparison of the first two most important variables selected by PCA (unsupervised learning technique) and PLS-DA shows the advantage of prior knowledge about observation membership in building classification model.



PLS-DA MODEL TUNING

Tuning the PLS-DA model allows for further improvement of sensitivity and selectivity. In the first stage the number of components required for the model is evaluated. Then the model is fine tuned to include only these original variables that are necessary to discriminate the classes.



CLASSIFICATION QUALITY EVALUATION

Evaluation of the classification quality may be done by comparing the group membership for every observation calculated with the model (predicted value) and the actual value (evaluated by the histopathologist). Another factor that should be taken into account is computational time.
As expected supervised learning performs better (especially PLS-DA technique) however in this case the longest time is required to complete the modeling.

CONCLUSIONS

Prostate cancer is one of the most serious threat that million of men face nowadays. Adequate medical treatment highly depends on proper recognition of the stage of disease. Traditional assessment based purely on histological cellular patterns is prone to human errors, even if it is done by very experienced specialist.
New system that includes additional information of elemental concentrations within prostate tissues may help to improve and simplify the classification.
Models based on unsupervised methods produce less accurate results created groups are characterized by the lack of pronounced borders, so proper classification of new cases seems to be impossible. On the other hand supervised methods, especially tuned PLS-DA tested here, delivers better results in terms of error-free grouping. The only disadvantage of this technique is long computational time.

REFERENCES

[1] R Core Team. R: A Language and Environment for Statistical Computing, 2017. URL http://www.r-project.org/.

[2] RStudio. RStudio: Integrated development environment for R, 2017. URL http://www.rstudio.com/.

[3] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. ISBN 978-0-387-98140-6. URL http://had.co.nz/ggplot2/book.

[4] Hadley Wickham and Romain Francois. *dplyr: A Grammar of Data Manipulation*, 2016. URL https://CRAN.R-project.org/package=dplyr. R package version 0.5.0.

[5] Sébastien Lê, Julie Josse, and François Husson. FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, 25 (1):1–18, 2008. doi: 10.18637/jss.v025.i01.

[6] Kim-Anh Le Cao, Florian Rohart, Ignacio Gonzalez, Sebastien Dejean with key contributors Benoit Gautier, Francois Bartolo, contributions from Pierre Monget, Jeff Coquery, FangZou Yao, and Benoit Liquet. *mixOmics: Omics Data Integration Project*, 2017. URL https://CRAN.R-project.org/package=mixOmics. R package version 6.1.2.

E-mail: slskb@nus.edu.sg Website: ssls.nus.edu.sg