Krajowa Konferencja Zastosowań Matematyki w Biologii i Medycynie

Sandomierz, 5th–9th September 2016

TRAINING OF CROSS-PLATFORM GENETIC SIGNATURES WITH APPLICATIONS TO THE CANCER GENOME ATLAS DATA

Przemyslaw Biecek[1,2]

[1]Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw
[2]Faculty of Mathematics and Information Science, Warsaw University of Technology
Przemyslaw.Biecek@gmail.com

ABSTRACT

The Cancer Genome Atlas is a rich source of genetic data that includes samples from 11,000 patients across 33 tumour types. It is also a collection of data from different platforms, among them profiles of DNA methylation, SNP and CNV variants, genes expressions measured with RNA-seq and proteins levels measured with RPPA arrays. Supplemented with rich clinical data this database creates opportunity for new interesting cross-platform genetic signatures.

In this talk we will compare three state of the art approaches for training of such genetic signatures, namely random forest, tree based gradient boosting and regularised logistic regression. Signatures are trained based on hundreds of thousands biomarkers. We will discuss challenges and opportunities behind each method and present an example application.

ACKNOWLEDGEMENTS

We would like to thank Maciej Wiznerowicz and his team from GreaterPoland Cancer Center and Maciej Żylicz and his team from International Institute of Molecular and Cell Biology in Warsaw for their comments and suggestions related to biological and medical grounds of cancer signatures.

REFERENCES

[1] Marcin Kosinski and Przemyslaw Biecek: *The family of R packages containing TCGA data*, MI2 group, 2015, https://github.com/mi2-warsaw/RTCGA.data.

[2] Przemyslaw Biecek, Ewa Szczurek, Martin Vingron, and Jerzy Tiuryn: *The R Package bgmm: Mixture Modeling with Uncertain Knowledge*, Journal of Statistical Software **47 (3)** (2012), 1–32, http://www.jstatsoft.org/v47/i03/.

[3] Andy Liaw and Matthew Wiener: *Classification and Regression by randomForest*, R News **2 (3)** (2002), 18–22.

[4] J.H. Friedman: *Greedy Function Approximation: A Gradient Boosting Machine,* Annals of Statistics **29 (5)** (2001), 1189–1232.

[5] Przemyslaw Biecek and Marcin Kosinski: *archivist: Tools for Storing, Restoring and Searching for R Objects*, CRAN R package version 1.5, 2015.

[6] R Core Team: *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, 2015, http://www.R-project.org/.

[7] Sage Bionetworks: *Synapse Contribute to the Cure* (2015), https://www.synapse.org/.

[8] National Cancer Institute: *The Cancer Genome Atlas* (2015), http://cancergenome.nih.gov.

[9] Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, and Alvis Brazma: *BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis*, vol. 21, Bioinformatics, 2005.