

Computation and Selection of Optimal Biomarker Combinations for Omics Data Using a ROC summary measure

Claudio Junior Salaroli and Maria del Carmen Pardo

Department of Statistics and O.R., Complutense University of Madrid

March 16, 2021

In *-omics* data, dimensionality is one of the main problems to deal with. Selecting and combining only the real predictive biomarkers allows to improve performances and interpretability of the results, omitting the noise caused by irrelevant features. Penalization techniques permit to achieve this aim, searching the feature selection space and the learning algorithm parameter space at the same time, returning a sparse solution. We propose PYE, the Penalized Youden Index Estimator, a method that combines Youden Index estimator and penalization techniques. PYE can select and combine biomarkers at the same time in a high-dimensional context, identifying also the optimal cut-off point. The performances of the proposed PYE classification framework were tested on different types of *-omics* datasets. Our method showed top results in both selection and combination, with a significant gain compared to some benchmark competitors.

Keywords: Penalized Youden index, Youden index, selection and combination of biomarkers, disease classification, high dimensionality.

References

- Yin, J, Tian, L.** (2013). Optimal linear combinations of multiple diagnostic biomarkers based on Youden index *Statistics in Medicine*, **33**:1426-1440.
- Horgan R.P., Kenny L.C.** (2011). 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics *The Obstetrician & Gynaecologist*, **13**(3):189-195.
- Saeys, Y., Inza, I., Larrañaga P..** (2007). A review of feature selection techniques in bioinformatics *Bioinformatics*, **23**(19):2507–2517.
- Hastie, T., Tibshirani R., Friedman, J.** (2008). The Elements of Statistical Learning. Data Mining, Inference, and Prediction *Springer Series in Statistics*.