More is Better: Ranking with Multiple Targets for Biomarker Discovery

Dragi Kocev, Ivica Slavkov, and Sašo Džeroski

Department of Knowledge Technologies, Jozef Stefan Institute,
Jamova 39, 1000 Ljubljana, Slovenia
{Dragi.Kocev,Ivica.Slavkov,Saso.Dzeroski}@ijs.si

The process of biomarker discovery is equivalent to the process of feature ranking and selection in machine learning. Each marker has a relevance measure assigned to it by a ranking algorithm. Typically, the ranking is produced with respect to a single target variable (e.g. outcome of a disease). But, the clinical data for a patient is much more complex and it contains multiple variables of interest.

Here, we address the problem of feature ranking in the context of multiple targets. In particular, we propose an extension of feature ranking with Random Forests (RFs) by enabling the method to handle multiple target variables simultaneously. The feature importance measure is calculated by randomly permuting the values of the features and measuring the out-of-bag (OOB) error estimates. The rationale is that if a feature is important for the target concepts it should have an increased error rate when its values are randomly permuted.

We apply the proposed method for feature ranking with multiple targets to Neuroblastoma microarray data associated with clinical data containing multiple variables of interest. We produce ranked lists of genes with respect to different (single) clinical parameters and compare these ranked lists with the one produced by considering multiple target variables simultaneously. We compare the ranked lists by using so-called average testing error curves (ATEs), which give us an estimate of the predictive performance of the highly ranked genes (markers). The results show an increase in the predictive performance of the highly ranked genes, when considering multiple target variables as compared to the ones from the ranked lists for each target variable individually. It is important to note that the same set of markers produced by the multiple target approach can be used for predicting the different clinical variables instead of having a different set of markers for each one.

In summary, we consider the process of biomarker discovery from a perspective of single vs. multiple target variables. The intuition behind using multiple target variables simultaneously comes from the usual complexity of the diseases under consideration (e.g. cancer) and the associated multi-variable patient clinical data. Our initial results show that the multiple target approach is beneficial as compared to the single target variable approach. The produced ranked list of biomarkers is more accurate, in terms of predictive performance, and it can be applied to each of the target variables separately.