

Hierarchical Multi-label Classification of ToxCast Datasets

Nina Jeliaskova nina@acad.bg, Vedrin Jeliaskov vedrin@acad.bg

Ideaconsult Ltd., Angel Kanchev Str 4, 1000 Sofia, Bulgaria

Abstract

Hierarchical multi-label classification is a variant of classification where instances may belong to multiple classes at the same time and these classes are organized in a hierarchy. We present the results of a feasibility study on the induction of hierarchical multi-label classification decision trees from the ToxCast datasets, with the objective to correlate in-vitro data with ToxRefDB in-vivo test results. A multi-label classification approach allows utilizing all the information available in ToxRefDB database simultaneously, thus exploiting potential relationships between various in-vivo test results. A tree or an acyclic graph hierarchy can be imposed over the classes providing means to group classes on e.g. study type, species or target site. The skewed distribution of ToxCast in-vitro data (the number of active entries does not exceed 10%) presents a challenge to standard approaches for feature selection and learning algorithms. We discuss suitable pre-processing techniques and the efficiency of model learning with different hierarchical constraints. The interpretability and relevance of the models in the context of ToxCast assays is explored. The predictive performance, model size and induction time are investigated in depth in the experiments section.