[P20] Interpretation of QSAR models learned from imbalanced data

Mariia Matveieva, Pavel Polishchuk

Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University Olomouc, Hněvotínská 1333/5, 779 00 Olomouc, Czech Republic

Interpretation of QSAR models can provide valuable information about structure-property relationship. The approach to QSAR model interpretation which estimates contribution of an arbitrary molecular fragment to the property modeled was developed earlier and applied to many classification and regression tasks.

HTS results in a huge amount of data which is a valuable source for mining of structure-activity relationships. However, data sets coming from HTS assays usually contain a small number of active compounds and, therefore, have high imbalance ratio. This frequently results in low predictivity of developed QSAR models and therefore should affect interpretation outcome.

In present study we investigated imbalanced classification problem in QSAR modeling and influence of class imbalance on model interpretability. We built models on imbalanced data sets to obtain baseline models. Further, models using multiple undersampling technique were built. It was shown that the latter models achieve higher predictive performance estimated by cross-validation.

This workflow was applied to modeling of anticancer activity based on the proprietary data set and several public data sets. The superiority of interpretation of models obtained using multiple undersampling over baseline models has been demonstrated. The automated approach to building models using multiple undersampling and interpretation of them was implemented in open-source SPCI software (https://github.com/DrrDom/spci).