The influence of training actives/inactives ratio on machine learning performance

Rafał Kurczab[1*], Sabina Smusz[1,2], Andrzej J. Bojarski[1]

[1]Department of Medicinal Chemistry, Institute of Pharmacology Polish Academy of Sciences, Kraków, 31-343, Poland
[2]Faculty of Chemistry, Jagiellonian University, Kraków, 30-060, Poland

*kurczab@if-pan.krakow.pl

In drug discovery, machine learning is widely used to classify molecules as active or inactive against a particular target. The vast majority of these methods (supervised learning) needs a training set of objects (molecules) to develop a decision rule that can be used to classify new entities (the test set) into one of the two mentioned classes [1].

A lot of studies, searching an optimal learning parameters and their impact on classification effectiveness were performed [2,3]. Unfortunately, there is no data showing the influence of actives/inactives ratio, used to model training, on the efficiency of new active compounds identification. Therefore, the main goal of this study was to examine the impact of changing the number of inactives in the training set with fixed amount of actives. For a given ratio, the inactives were randomly selected from ZINC database (10-times to prevent an overestimations error). This concept was verified on three different protein targets (i.e. 5-HT_{1A}, HIV-1 protease and matrix metalloproteinase) and a set of algorithms (SMO, Naïve Bayes, Ibk, J48 and Random Forest) implemented in WEKA package [4]. To compounds representation, two types of molecular fingerprints were used (MACCS and hashed fingerprint), to determine their possible impact on machine learning performance.

Acknowledgements

The study was supported by a grant PRELUDIUM 2011/03/N/NZ2/02478 financed by the National Science Centre.

References

[1]. Melville JL, Burke EK, Hirst JD: **Machine learning in virtual screening.** *Comb Chem & High Thr Scr* 2009, **12**:332–343.

[2]. Ma XH, Wang R, Yang SY, Li R, Xue Y, Wei YC, Low BC, Chen YZ: **Evaluation of virtual screening performance of support vector machines trained by sparsely distributed active compounds**. *J Chem Inf Mod* 2008, **48**:1227-37.

[3]. Plewczynski D, Spieser SH, Koch U: **Assessing different classification methods for virtual screening.** *J Chem Inf Mod* 2006, **46**:1098-106.

[4]. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH: **The WEKA data mining software: an update.** 2009 *SIGKDD Explorations*, **11(1)**:10-18.