The influence of training set size on machine learning performance

Rafał Kurczab, Sabina Smusz, Andrzej J. Bojarski

Department of Medicinal Chemistry, Institute of Pharmacology Polish Academy of Sciences, Smętna 12 Street, 31-343 Kraków

e-mail: kurczab@if-pan.krakow.pl

Introduction

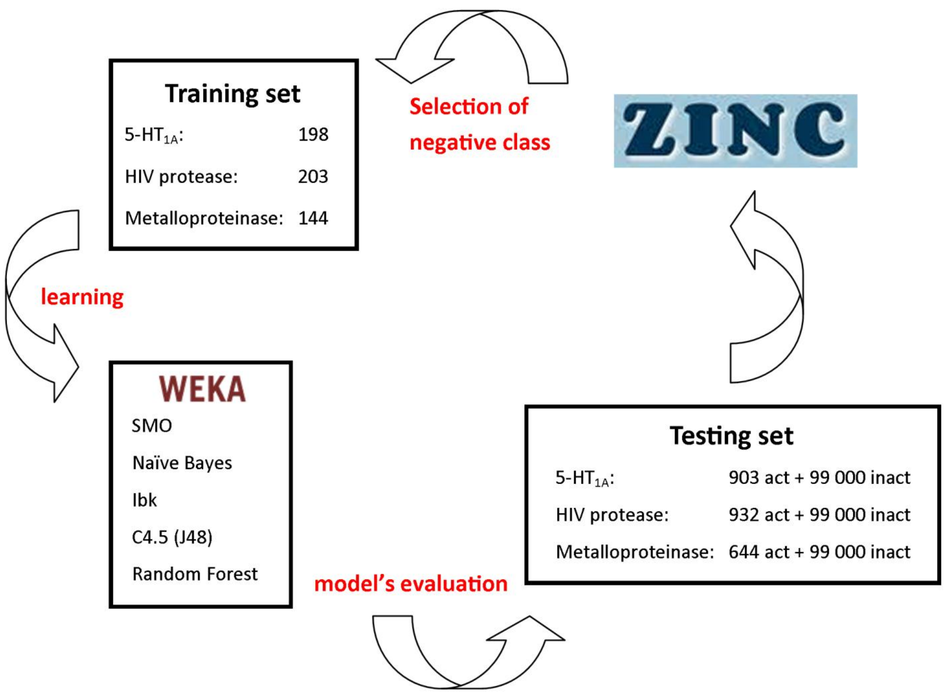
In drug discovery, the machine learning is widely used to classify molecules as actives or inactives against a particular target. The vast majority of those methods (supervised learning) needs a training set of objects (molecules) to develop a decision rule that can be used to classify new molecules (the test set) into one of the two activity classes [1]. A lot of studies looking for an optimal learning parameters and their impact on classification effectiveness were performed [2,3]. Unfortunately, there is no data showing the influence of actives to inactives ratio, used to model training, on the efficiency of new active compounds identification.

Herein, we present the results of the increase the number of inactives used in the training of the some machine learning algorithms. The effects of this changes were measured on the virtual screening experiment.

Materials and Methods

The general computational procedure applied in this study is presented on the scheme below. The MDDR database was used to extract the structures of known active compounds for three different protein targets (i.e. 5-HT_{1A} agonists, HIV protease inhibitors and metalloproteinase inhibitors). They were next divided on the training and testing sets, where consisted an positive learning examples. The negative instances were chosen from the ZINC database (assumed as inactives) by the random selection. Only one set was established for evaluation test, common for all targets, whereas for training set it was selected 10-times, for a particular actives to inactives ratio. The size of active class was fixed, therefor to change the ratio, the amount of negatives were varied (form 100 to 2000 with the step equal 100).

The experiment were performed for two different molecular fingerprints (MACCS and hashed FP) obtained by the use of PaDEL-Descriptor software [4], and a set of machine learning algorithms (SMO, Naïve Bayes, Ibk, J48 and Random Forest) implemented in WEKA package [5]. To monitor the changes in the method's performance the three parameters were used: recall, precision and Mathews Correlation Coefficient.



Scheme 1. Computational workflow applied in this study.

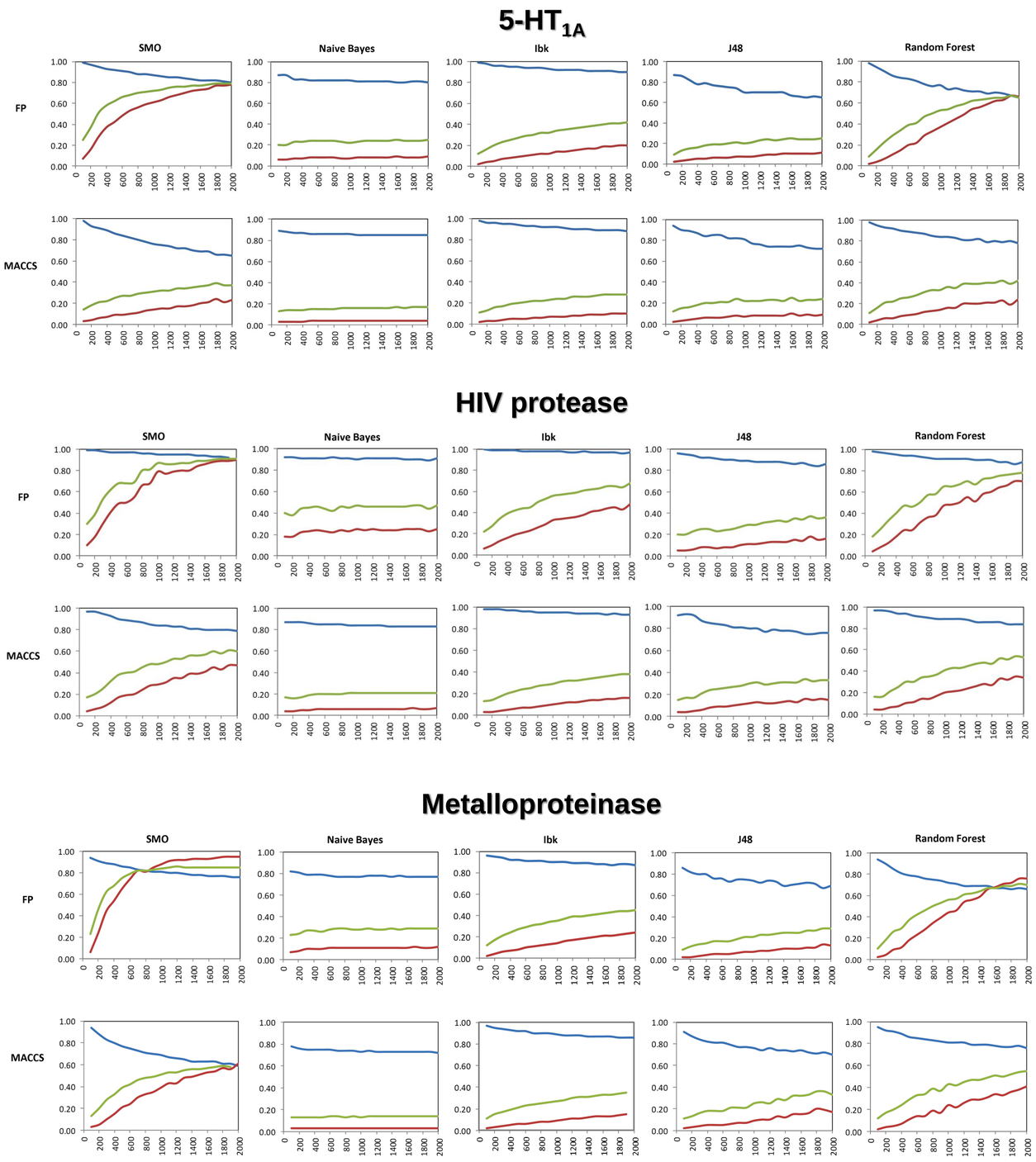


Figure 1. The graphs showing the influence of inactive set size changes on the performance of machine learning methods. The blue line denote the changes of recall value, whereas the red and green the precision and MCC, respectively.

Conclusions

The performance of machine learning methods depends on the actives to inactives ratio, used to training the prediction models. Moreover, it seems that this step can be used as a boosting-like method of machine learning method's improvement.

In general, increasing the number of negative instances, when the amount of positives is constant, caused the decreasing the hit recall, but increasing the prediction accuracy (precision). The observation of global performance parameter (MCC) showed that the prediction ability of the methods are growing up. An exception is the Naive Bayes algorithm for which no significant changes were observed. This provide some evidence of their independence of the training set perturbation and variations.

The results are consistent between targets and fingerprints, however fingerprint with lower length (MACCS) might not be a proper to improve the training models.

Results

The results for all molecular targets and combinations of fingerprint and machine learning method were shown in figure 1. The single plots present the dependence between averaged (for all iterations) value of given performance measure and an inactive set size, obtained for a particular method.

In almost all cases, the decrease of the recall and increase of the precision and MCC values simultaneously, when the size of negative instances was raised is visible. A completely different behavior shown the Naïve Bayes algorithm, which in all cases was not sensitive on any changes of the set size.

Taking the dynamic of recall decrease (minimal) and increase of precision (maximal), the SMO and Random Forest methods showed the best rate. Additionally, the Ibk algorithm seems to produce the lowest decrease of recall, when the set size is increase.

The overall tendency did not changed when the MACCS fingerprint was used. In general, it causes deterioration of the obtained results.

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] Yap CW: **PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints.** *J Comp Chem* 2011, **32(7)**:1466-1474

[5] 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.

Acknowledgments

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