Meta-learning as an improvement of machine learning methods performance in virtual screening

Sabina Smusz[a,b], Rafał Kurczab[a], Andrzej J. Bojarski[a]

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

[b] *Department of Crystallochemistry of Drugs, Faculty of Chemistry, Jagiellonian University, 3 Ingardena Street, 30-060 Kraków*

Introduction

A great number of computational approaches have been developed in order to facilitate and improve the process of drug design. One of the most popular methodologies is virtual screening, that enables a selection of drug candidates out of large libraries of chemical compounds. Recently, many applications of machine learning methods in this strategy have been reported. Their main goal is to discover relationships between different features of existing data and use them for classification of unknown instances.[1]

Meta-learning

Meta-learning approach is often described as „learning to learn". It is connected with accumulating experience from analyzing a base-learning system performance. The typical tasks of meta-classifiers include:

- evaluation and comparison of learning methods,
- measurement the benefits of base-learning on learning on the meta-level,
- selection of useful strategies and discarding ineffective ones.

There are many different meta-algorithms but bagging and boosting are among the most popular and the most frequently used ones.[2]

Experimental part

Two training sets containing different number of 5-HT_{1A} antagonists (train1 < train2) and one test set were constructed. Active molecules were taken from the MDDR database and the inactive ones were randomly selected from the ZINC database. Then, for each structure eight different types of fingerprints were generated using PaDEL-Descriptor.[3] Machine learning methods were evaluated with the use of tools offered by WEKA package.[4] Performance of four meta-classifiers (MultiboostAB, Decorate, FilteredClassifier and Bagging) was tested in combination with three different base-learners (J48, RandomForest and NaiveBayes with recall, precision and MCC values as evaluating parameters. Their dependence on the type of fingerprint and the number of actives in the training set for selected methods was examined.

Results and Discussion

Values of all evaluating parameters are dependent on the type of fingerprint and the number of actives present in the training set (Figures 1–3). The dependence of recall and MCC on those factors are quite strong, but precision values are relatively high (~0.9) regardless of classification conditions. The higher number of actives in the training set, the higher values of recall and MCC (precision slightly falls with increasing number of actives in the training data).

Using meta-classifiers is usually connected with better performance in comparison with the peformance of base-learners alone (Figures 4–6). FilteredClassifier leads to lower values of evaluating parameters but the rest of seleted meta-algorithms provides (in most cases) the improvement of classification effectiveness. J48 is the classifier which is the most prone to the influence of meta-learning (using J48 in combination with MultiBoostAB, Decorate and Bagging leads to ~5% uplift in recall values, about 5–10% in precision values and ~10% when it comes to the values of MCC). The scale of improvement depends on classification conditions and is the greatest (almost 15% in MCC for train set containing more actives – Figure 6) when Extended Fingerprint is used as molecules representation and the lowest (~1–2% concerning MCC) when compounds are defined by Estate or Substructure Fingerprint. It is very difficult to choose the best meta-strategy, but on average using Decorate was connected with the highest improvement of base-classifier performance.

Acknowledgements

The study was partly supported by a grant PNRF-103-AI-1/07 from Norway through the Norwegian Financial Mechanism within the Polish-Norwegian Research Fund.

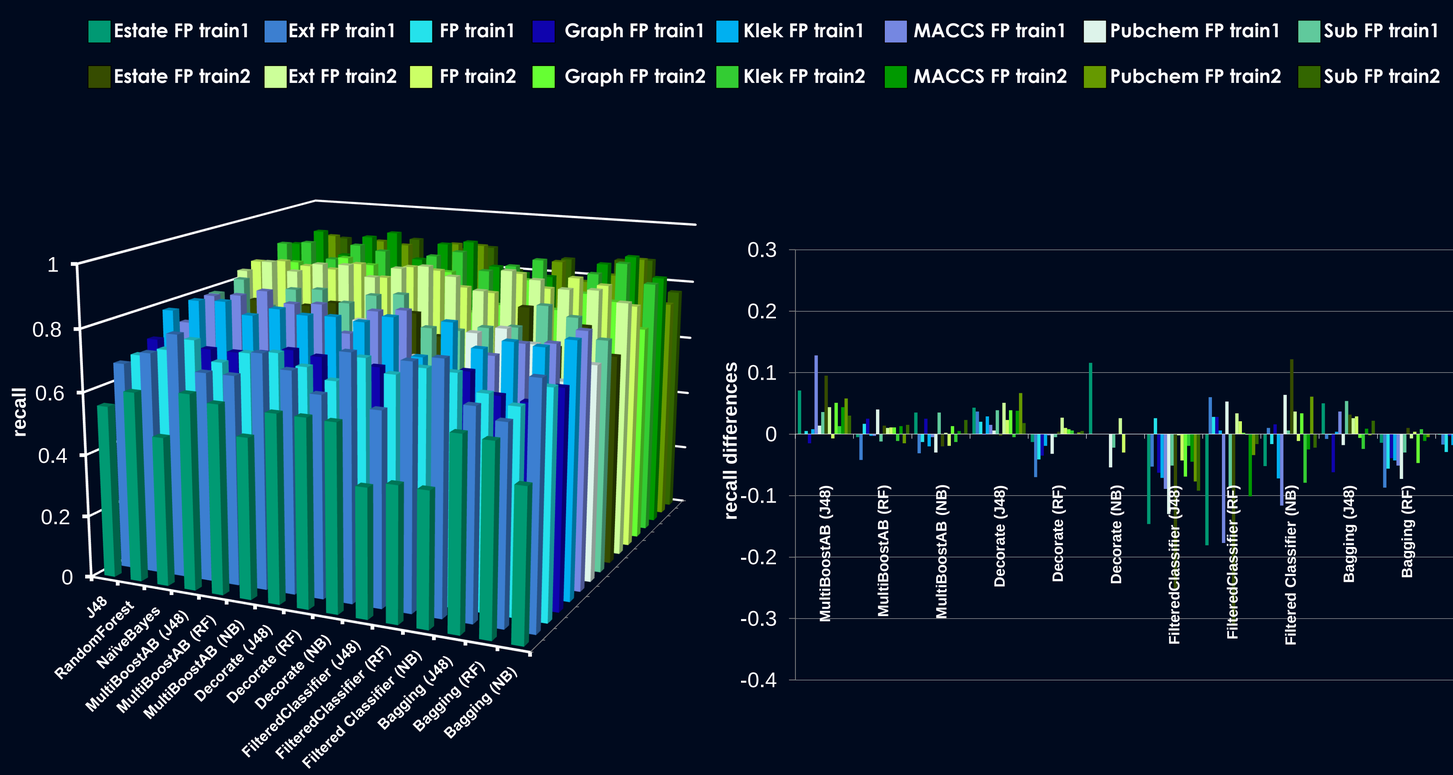

Figure 1. Recall values obtained for base- and meta-classifiers.

Figure 4. Differences in recall values obtained for meta- and base-classifiers.

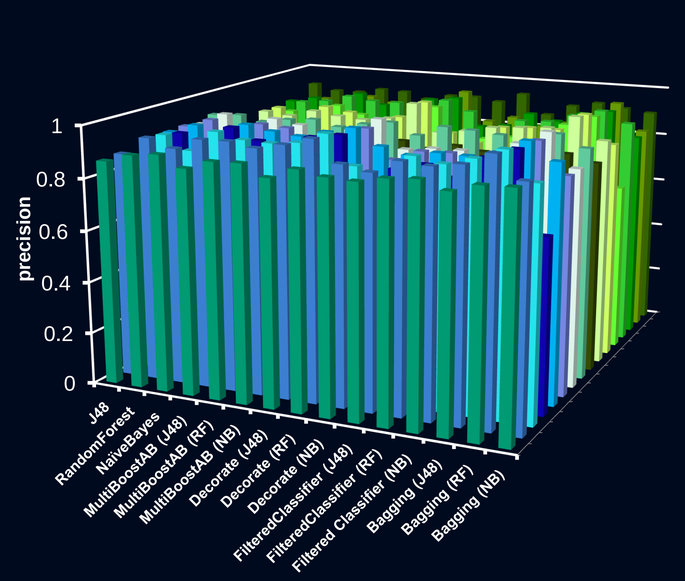

Figure 2. Precision values obtained for base- and meta-classifiers.

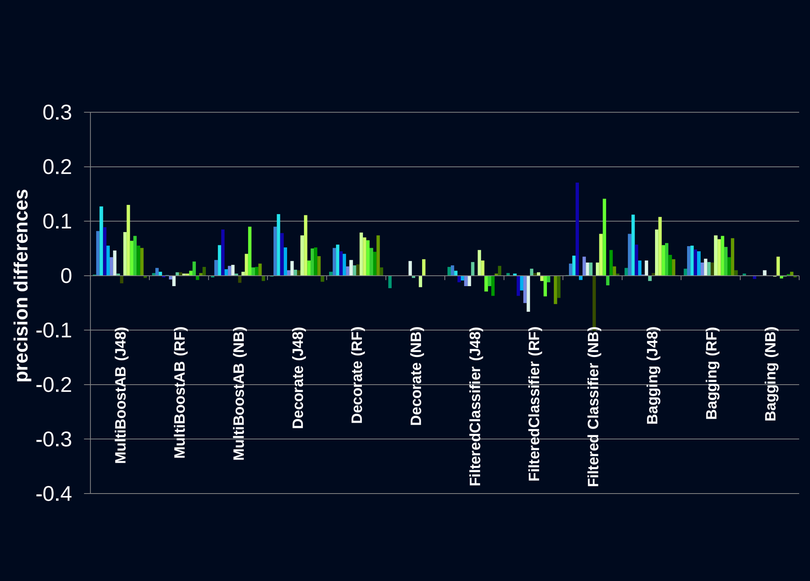

Figure 5. Differences in precision values obtained for meta- and base-classifiers.

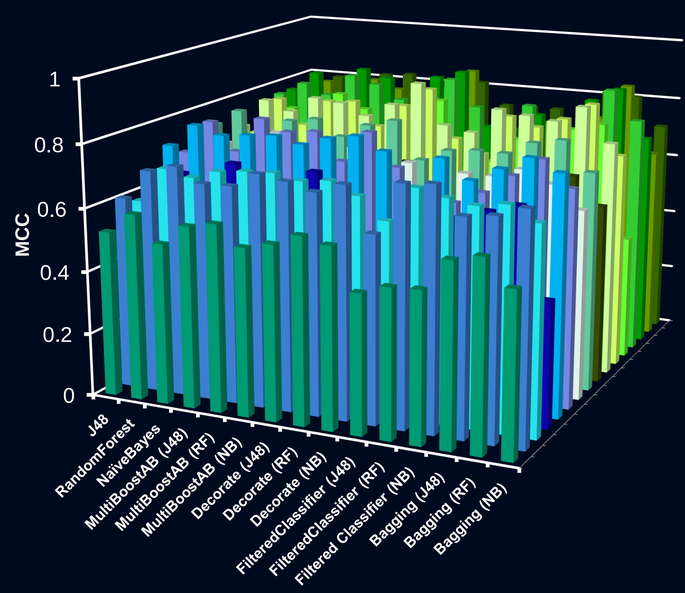

Figure 3. MCC values obtained for base- and meta-classifiers.

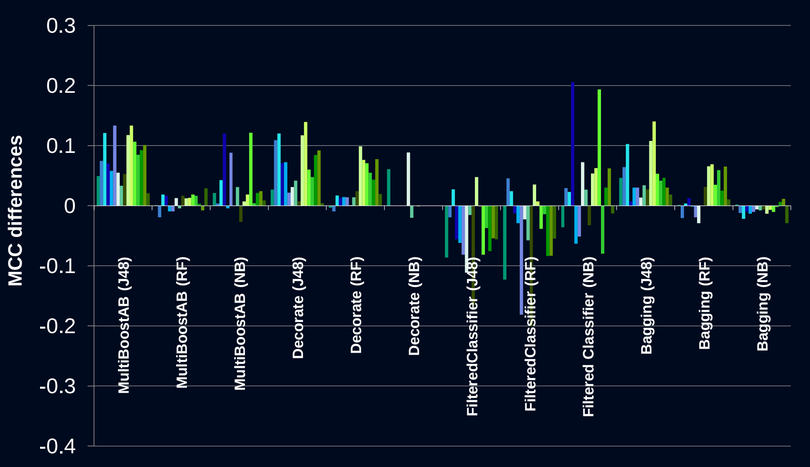

Figure 6. Differences in MCC values obtained for meta- and base-classifiers.

References

[1] Melville, J.L., Burke, E.K., Hirst, J.D. *Comb. Chem. High Throughput Screen.* 12 (2009) 332.

[2] Schmidhuber, J; Zhao, J.; Wiering, M. *Technical Report IDSIA* (1996) 69.

[3] Yap, C.W. *J. Comput. Chem*. 32 (2011) 1466.

[4] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H. *SIGKDD Explorations* 11 (2009) 10.