[P24] Challenge of a priori classification of organic reaction yields and duration

Grzegorz Skoraczyński[1], Piotr Dittwald[2], Bartosz Grzybowski[2], Błażej Miasojedow[1], Anna Gambin[1]

[1] Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Banacha 2, 02-097 Warszawa, Poland
[2] Institute of Organic Chemistry of the Polish Academy of Sciences, Kasprzaka 44/52, 01-224 Warszawa, Poland

Machine learning methods turned to be very efficient and popular in various applications where the goal is to predict the answer from multidimensional vector of features. The huge amount of data is processed to create an appropriate model that can be used for classification purpose.

Here we study the problem of automatic prediction of organic reaction yield and duration. The space of features is chosen to be the widely used space of reactions descriptors and so-called fingerprints used for numerical description of the reaction equation. Descriptors describe features of most important molecules participating in reaction like number NH or OH groups or number of hydrogen bond donors. Fingerprints describe changes of molecules which are going through reaction. They were generated using rdkit package [1], fingerprints were also processed using method proposed by Schneider et al. [2]. We analyzed databases with over one million of unique entries, which describe reactions their yields and duration. After removing duplicates we obtained datasets with over 450000 entries, which were further analyzed.

We used several methods for binary classification like SVM [6], kNN, RF, logistic regression, but random forest classifier [3] performed best. The obtained classification accuracies were rather low: ca. 65% for reaction yield and ca. 75% for reaction duration. Moreover this pessimistic outcome cannot be significantly improved. We provide the rigorous proof of this statement by approximating the lower bounds for error of Bayes classifier using method proposed by Berisha et al. [4]. This method consists in calculating Friedman-Rafsky test statistic [5], which describes relation between two classes, which we are classifying into and then approximating Bayes errors. Our outcomes were twofold as we focused both on the lower and upper bound for classification accuracy. Obtained results were at levels of 23% and 18% for reaction yield and duration, respectively.

These results show that the performance of the machine learning methods cannot be significantly improved neither by changing the structural features used for model training, nor by increasing the number of features unless the new better performing descriptors would be applied. Our analysis showed a high complexity of the initial problem, which does not guarantee its full solution even when much larger data would be gathered. Nonetheless, have also shown that purely computational approach based on large organic synthesis database can support to some degree the decision in the laboratory routine.

Bibliography:

[1] Rdkit: Open-source chemoinformatics. [Online; accessed 11-April-2013].
[2] Nadine Schneider, Daniel M. Lowe, Roger A. Sayle, and Gregory A. Landrum. J. Chem. Inf. Model., 55(1):39–53, 2015
[3] Leo Breiman. Random forests. Machine learning, 45(1):5–32, 2001.
[4] Visar Berisha, Alan Wisler, Alfred O. IEEE Transactions on Signal Processing, 64(3):580–591, 2016.
[5] Jerome H. Friedman and Lawrence C. Rafsky. Ann. Statist., 7(4):697–717, 1979.
[6] Corinna Cortes and Vladimir Vapnik. Machine learning, 20(3):273–297, 1995.