[L5] The best practices for multilearning in (Big) data analysis

Igor V. Tetko

Institute of Structural Biology, Helmholtz Zentrum München – German Research Center for Environmental Health (GmbH), Ingolstaedter Landstrasse 1, 85764 Neuherberg, Germany.

Despite the increasing volume of data, their amount remains very small compared to the virtual chemical space of possible chemical structures. Therefore, there is a strong interest in simultaneous analysis of different ADMET and biological properties of molecules, which are frequently strongly correlated one with another. Apart from the increase of the sheer volume of data, such joint data analysis can increase the accuracy of models by exploiting their common representation and identifying common features between individual properties. The majority of multilearning approaches are developed for individual machine learning methods. Several examples of them for both linear and non-linear methods will be presented. Contrary to such methods, the feature nets can be applied to virtually any machine learning method.[1] Following the overview of the existing approaches we will explain when their application is expected to be beneficial following a comprehensive study of Xu et al [2]. The presentation will be completed with an overview of the multi-learning approaches implemented within the On-line Chemical Database and Modeling Environment (OCHEM).[3]

Acknowledgement

The project leading to this report has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 676434, "Big Data in Chemistry" http://bigchem.eu. The article reflects only the author's view and neither the European Commission nor the Research Executive Agency (REA) are responsible for any use that may be made of the information it contains.

Bibliography:

1. Varnek, A.; Gaudin, C.; Marcou, G.; Baskin, I.; Pandey, A.K.; Tetko, I.V. Inductive transfer of knowledge: Application of multi-task learning and feature net approaches to model tissue-air partition coefficients. J. Chem. Inf. Model. 2009, 49, 133-144.
2. Xu, Y.; Ma, J.; Liaw, A.; Sheridan, R.P.; Svetnik, V. Demystifying multitask deep neural networks for quantitative structure-activity relationships. J. Chem. Inf. Model. 2017, 57, 2490-2504.
3. Sushko, I.; Novotarskyi, S.; Korner, R.; Pandey, A.K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V.V.; Tanchuk, V.Y., et al. Online chemical modeling environment (ochem): Web platform for data storage, model development and publishing of chemical information. J. Comput. Aided. Mol. Des. 2011, 25, 533-554.