Filling data gaps of sustainability related properties based on molecular descriptors and advanced data mining methods

Background/ Motivation

The escalating production of existing chemicals and the continuing development of new chemical substances lead to increased exposure to various chemical molecules. To prevent or minimize the harmful impact on human health and the environment, the safety of the chemicals should be evaluated prior to their manufacturing and use. However, the safety assessment requires the availability of information on various molecular properties like toxicity, persistency, etc. which requires experimental testing. The testing has traditionally been performed on living organisms which, considering the continuously increasing numbers of the new molecules, is not feasible neither from economical nor ethical points of view. To minimize the amount of experimental testing development of reliable *in vitro* and *in silico* methods is required.

Various data mining techniques have recently been developed with the aim to find critical patterns in already available molecular data and utilize them to predict properties of existing or newly manufactured chemicals with missing experimental values. The methods are typically successful in forecasting properties where large amount of measured data exist. However, holistic assessment requires prediction of variety of properties to evaluate impact of the chemicals to people and environment more thoroughly, but the amount of experimental data for some properties of the chemicals is quite limited (e.g., acute dermal toxicity, permissible exposure levels, acidification potential, ozone depletion potential etc.).

Prior knowledge incorporation could improve the prediction. The prior knowledge can refer to the estimated functional relationship between target and predictor variables (e.g., hazard property and molecular structure) or input/ output information [1]. Such kind of knowledge, obtained by data mining techniques, heuristics or generated from theoretical observations or data provided by experts in the field, have been proven to increase the performance of data mining prediction models, if they are properly incorporated into it.

Variable constraints, predictors significance, correlation between predictors, information about the input data such as noise, outliers, way of addressing missing data can be the possible areas of knowledge exploration and generation with the aim to be further utilized in the prediction process. The knowledge can be incorporated in form of additional constraint functions or applied to choose or specify data mining model parameters.

Objective:

The overall objective of the project is to incorporate the knowledge to the prediction model and evaluate the influence of the knowledge incorporation on the prediction result.

This project includes the following steps:

- Investigate/develop methods for incorporating prior knowledge to selected data mining techniques to create a hybrid property prediction model.

- Apply standalone data mining tools and compare with the hybrid approach to investigate the benefits of combining existing knowledge with black box data mining models.

Tools/Databases:

Data mining: Python & python packages: Spacy, Gensim, Pattern, Pandas, Scikit-learn, NumPy, Rdkit etc. Other: GATE, RapidMiner etc.

Prerequisites:

The applicant should have a background in chemistry or chemical engineering with interest or previous experience in data science methods and applications or, vice versa, a background in computer science with interest or previous experience in chemistry or chemical engineering applications.

Supervisors:

Gulnara Shavalieva and Stavros Papadokonstantakis

Chalmers University of Technology, Division of Energy Technology

gulnara.shavalieva@chalmers.se

stavros.papadokonstantakis@chalmers.se

References:

[1] S. Chen, C. Gao, and P. Zhang, "Enhancing Transparency of Black-box Soft-margin SVM by Integrating Data-based Prior Information," 2017.