ABSTRACT

Sustainability assessment using *in silico* methods

Gulnara Shavalieva and Stavros Papadokonstantakis

Chalmers University of Technology, Department of Space, Earth and Environment, Division of Energy Technology, SE-412 96 Gothenburg, Sweden

There were around 95 million tons of chemicals produced in EU during the year 2016. Most of them are substances hazardous to health and environment [1]. Safety assessment of such chemicals require molecular property information that has been traditionally obtained by the means of animal testing. Current EU legislation strongly encourages to replace cruel testing by *in vitro* or *in silico* methods [2]. Therefore, development of reliable predicting methods is necessary.

Various data mining techniques has been recently developed with the aim to find critical patterns in already available data and utilize them to predict properties of newly manufactured and existing 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 molecular properties to evaluate impact of the chemicals to people and environment more thoroughly but amount of experimental data for some properties is quite limited (e.g., acute dermal toxicity, permissible exposure levels, acidification potential, ozone depletion potential etc.). Moreover, computer-aided molecular design (CAMD) studies come up with novel molecules for which no measurements exist at all.

As an example, aquatic toxicity ($LC50_{aq}$) and a database containing aliphatic molecules consisting of sixteen functional groups relevant to post – combustion CO_2 capture [3] was used to train (*kNN*) algorithms performance in terms of predictive capacity when the experimental data is scarce and imbalanced. The model were tested using 101 CO_2 capture amine solvents identified by means of CAMD and described in [4]. Coefficient of determination ($Q^{2)}$ for the toxicity $LC50_{aq}$ property prediction of the solvents is computed to be 0.38 [3].

Prior knowledge incorporation could improve the prediction. Expert knowledge or rules obtained by data mining techniques, heuristics or generated from theoretical observations have been proven to increase the performance of data mining models, if they are properly incorporated into it. The knowledge can refer to the estimated functional relationship between target and predictor variables (e.g., hazard property and molecular structure) or input/ output data [5]. 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 created knowledge can be incorporated in form of additional constraint functions or applied to choose or specify data mining model parameters. In case of the property prediction for solvents presented above, knowledge about the main structural contributor to the toxicity of the molecule could assist at selecting the most similar compounds in terms of toxicity. Significance of the contributors could also be used to adjust distance calculation to evaluate similarity between the molecules.

Future research should further develop and confirm the knowledge incorporation ways and benefits.

References:

[1] Chemicals production and consumption statistics. [online]. [Accessed 19 October 2018]. Available from: https://ec.europa.eu/eurostat/statistics-explained/index.php/Chemicals_production_and_consumption_statistics#Further_Eurostat_information

[2] Alternatives to animal testing and safety assessment of chemicals - EU Science Hub - European Commission. EU Science Hub [online]. 27 July 2017. [Accessed 19 October 2018]. Available from: https://ec.europa.eu/jrc/en/research-topic/alternatives-animal-testing-and-safety-assessment-chemicals

[3] Shavalieva, G, Papadopoulos, A.I., Badr, S., Seferlis, P., Papadokonstantakis, S.. Sustainability assessment using local lazy learning: The case of post-combustion CO2 capture solvents. In: Mario R. Eden, Marianthi Ierapetritou and Gavin P. Towler (Editors) Proceedings of the 13th International Symposium on Process Systems Engineering – PSE 2018 July 1-5, 2018, San Diego, California, USA © 2018 Elsevier

[4] Papadopoulos, A.I., Badr, S., Chremos, A., Forte, E.,Zarogiannis, T., Seferlis, P., Papadokonstantakis,S., Galindo,A., Jackson, G., and Adjiman, C..S., 2016b, Computer-aided molecular design and selection of CO2 capture solvents based on thermodynamics, reactivity and sustainability, Mol. Syst. Des. Eng.,1.

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