Enhancing Industrial Data Analysis through Machine Learning-based Classification of Petrochemical Datasets

Rastislav Fáber[1], Karol Ľubušký[2], Martin Mojto[1], Radoslav Paulen[1]

[1]*Faculty of Chemical and Food Technology, Slovak University of Technology in Bratislava, Bratislava, Slovakia*
[2]*Slovnaft, a.s., Bratislava, Slovakia*

e-mail: rastislav.faber@stuba.sk

Key words: Machine Learning, Data Classification, Alkylation Process, Analytics, Industry 4.0

Incorporating data analytics and machine learning (ML) algorithms into industrial decision making has proven to be a promising way to boost production efficiency. By utilizing ML algorithms to classify historical measurements from online sensors and laboratory analyses, it is possible to provide an operation guideline that was previously unavailable. We apply rigorous data treatment to prepare the raw data for ML-based classifier design. This process includes data cleaning, data standardization, data averaging, variable removal (based on linear dependency analysis), and distant outlier detection; to ensure the quality and reliability of available data. Selection of a suitable classifier model depends on the complexity of an industrial process, the level of its automation (implementation effort) and the ability to handle data outliers. We employ Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for initial ground-truth labeling, after which we utilize well understood ML algorithms; k-Means, k-Nearest Neighbors (k-NN), Support Vector Machine (SVM) and SVM with time difference, to engineer a framework for real-time classification. Accurate categorization of measurements is crucial for identifying slight deviations from real values that could impact the quality of the final product. Moreover, the complexity of the data plays a significant role in the performance of ML algorithms. With precise categorization of real-time data, the need for human intervention in process control can be minimized. To evaluate the performance of the designed classifiers, we compare their classification accuracy against the aforementioned synthetic ground truth labels. This comparison is carried out on a testing dataset that was not used during the framework design. Overall, our results demonstrate that the ML-based classifiers achieve comparable results in real-time classification. The most accurate classifier was the SVM model which uses not only absolute data, but also their time differences, which achieved the highest anomaly detection, 82 %.

Acknowledgments: This research is funded by the Slovak Research and Development Agency under the project APVV-21-0019, and by the Scientific Grant Agency of the Slovak Republic under the grants VEGA 1/0297/22 and VEGA 1/0691/21, and by the European Commission under the grant no. 101079342 (Fostering Opportunities Towards Slovak Excellence in Advanced Control for Smart Industries).