MODELLING OF CRYSTALLIZATION CONDITIONS FOR ORGANIC SYNTHESIS PRODUCT PURIFICATION USING DEEP LEARNING

Mantas Vaškevičius, Jurgita Kapočiūtė-Dzikienė
Informatics Faculty, Vytautas Magnus University

Introduction

Crystallization is used in the chemistry laboratory as a purification technique for solids. It is one of the fundamental procedures in the laboratory that is based on the principles of solubility. An important feature of crystallization is selection of an appropriate solvent. The aim of the paper is to formulate reasonable approaches based on modern machine learning algorithms to predict appropriate solvents for purification of synthesis mixtures using crystallization.

Data and methods

Daniel Mark Lowe's and NextMove's open-source collection of chemical reactions extracted from US patents issued from 1976–2016 has been used to create the dataset for testing of machine learning algorithms. The original dataset is comprised of 3.7 million reactions and synthesis procedures. We have made use of structured synthesis procedures to extract the solvent names used in the crystallization step of the syntheses. We have used two datasets in total; the additional dataset has supplementary input information that describes which solvents were in the mixture before crystallisation and may influence the prediction accuracy.

We have selected two vectorization methods, namely extended-connectivity fingerprints (ECFP) and ECFP autoencoders. ECFPs denote an absence or existence of specific substructures by scanning atom neighbors. ECFP autoencoders can be utilized for dimensionality reduction for sparse matrices, such as extended-connectivity fingerprints. We also test two types of neural networks (feed-forward neural network (FFNN) and long short-term memory (LSTM)).

Results

The results are presented in two figures: with mixture information (fig. 1) and without (fig. 2). ECFP and ECFP + E notations decribe the vectorization method, while numbers in the parentheses denote the lenght of the ECFP vector. Feed-forward and LSTM neural networks have been used to train multi-label classifiers. The most accurate optimized model can predict solvent labels with an accuracy of 0.87 ± 0.004 (using 1024 unit long ECFP + E vectorization with LSTM neural networks).

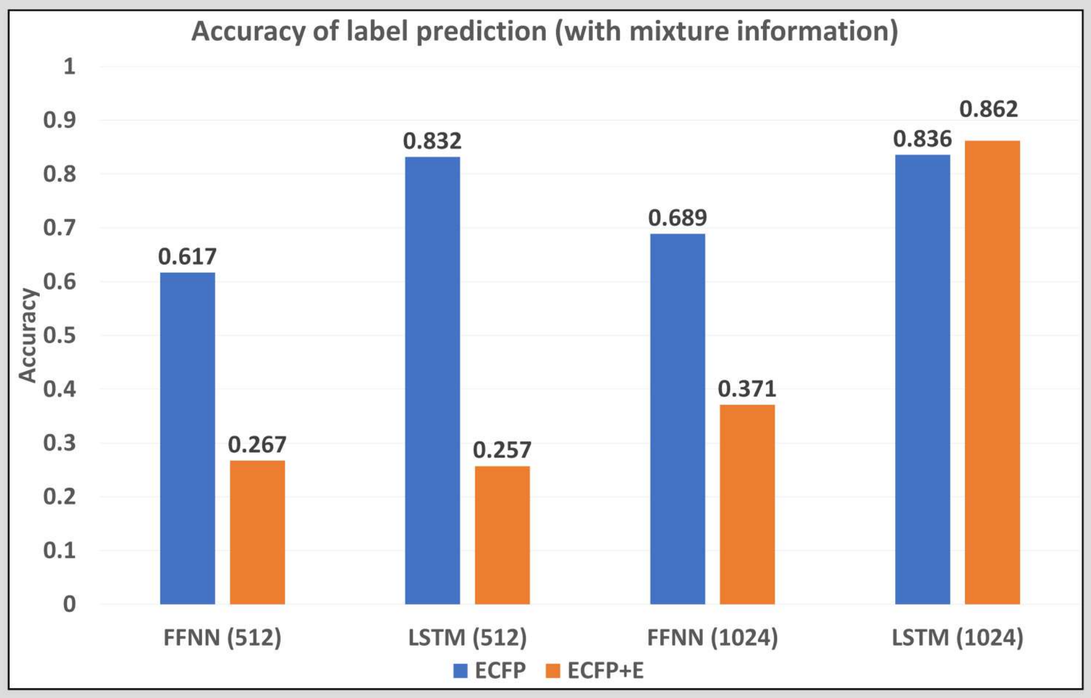

Figure 1. Accuracy values of label prediction with mixture information

Conclusions

In general, vectorization with autoencoders leads to lower performance, however, when used with longer vectors and LSTM neural networks, they are able to produce more accurate results. Research results may be used to accelerate R&D processes in the laboratories. Additionally, a method for modeling of reaction mixture crystallization conditions invariant to the reaction type has not been previously demonstrated.

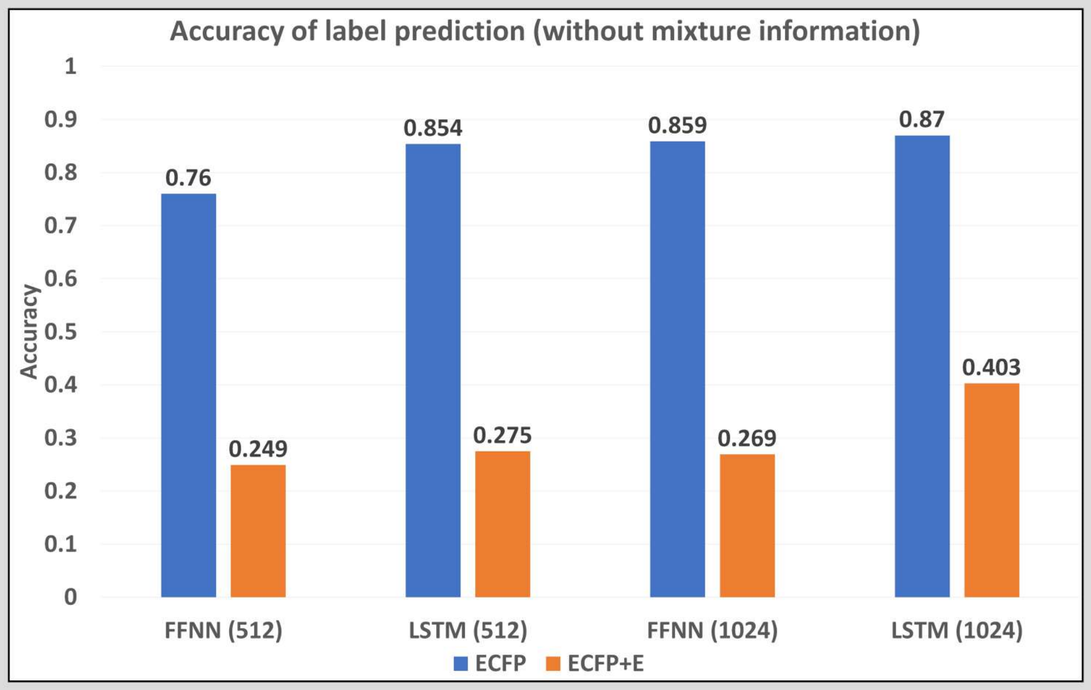

Figure 1. Accuracy values of label prediction without mixture information