PharmSci 275 Final Report

Prediction of small-molecule compound solubility in organic solvents by machine learning algorithms

authors: Zhuyifan Ye and Defang Ouyang

Aakankschit Nandkeolyar

March 2022

0.1 Paper Summary

This paper aims to utilize machine learning methods to build quantitative structure property relationship (QSPR) models.The main aim of these QSPR models is to be able to predict log of the solubility in various organic solvents based on structural properties of compounds based on their SMILE strings and temperature. To input structural properties of compounds into the models, an Extended-Connectivity fingerprint (ECFPs) was generated using the RDKit package in python. The data-set consisted of 5081 data points of solubility collected at different temperatures, it covered 266 compounds and 123 organic solvents. The models were created in two different training and testing methods-the first involved using a hold out method with a fifth of the data being used as a validation set and the other involved using a 5-fold cross validation procedure that randomly split the data into a 80% training data set and 20% testing data-set 5 times. A number of different machine learning models were utilized in this study these are listed in Table 1. These methods were evaluated using metrics from table 2.

Table 1: List of Machine Learning Methods

Model Names	**Type**	**Acronym**
Deep Neural Networks	Non-Linear	DNN
Support Vector Machines	Non-Linear	SVM
Gradient Boosted Method Decision Trees	Non-Linear	lightGBM
Partial Least Squares	Linear	PLS
k-nearest neighbors	Non-Linear	kNN
Ridge Regression	Linear	RR

All of the models were implemented in python using the sci-kit learn package except the lightGBM method which was implemented using Microsoft's lightGBM package. The hyper-parameters were optimized for the validation after the model was trained using the training data-set. It was found that the lightGB performed best compared to the other methods. Additionally, it was also found that non-linear methods outperformed the linear methods. The lightGBM model was chosen as the best model due to its ability to reduce bias and increase generality and essentially avoid over-fitting. Additionally it also allowed for further simplification of features by bundling mutually exclusive features and improving model robustness. Overall the lightGBM had a MAE of 0.2, MSE of 0.16 and an R^2 of 0.91.

Table 2: List of Evaluation Metrics

Parameter	**Formula**
Mean Absolute Deviation (MAE)	$\frac{\sum_{i=1}^{n} \lvert\hat{y}_i - y_i\rvert}{n}$ (1)
Mean Squared Error (MSE)	$\frac{\sum_{i=1}^{n} \lvert\hat{y}_i - y_i\rvert^2}{n}$ (2)
R-squared	$\frac{\sum_{i=1}^{n} \lvert\hat{y}_i - y_i\rvert^2}{\sum_{i=1}^{n} \lvert y_i - \bar{y_i}\rvert^2}$ (3)

0.2 Strengths

The main strengths of the paper lies in the use of lightGBM method, its ability to simplify the feature space for the model by bundling mutually exclusive features to improve the robustness of the models. It also reduces the number of features for a high dimensional sparse data, thus increasing the computational speed. It also has good performance, less over-fitting, higher efficiency and the ability to handle larger amounts of data. Its learning algorithm is histogram based which allows it to find the best partition values in features. These are important factors since they increase the generalizability of the model, which makes it more robust when making predictions.

The study also demonstrated that these machine learning models far out perform empirical models in making predictions associated with solubility. It also demonstrated that there exists a relationship between molecular finger prints that represent the structure of molecules and the prediction of solubility, it showed that multiple key structural aspects contributed to solubility prediction. The paper also successfully showed that prediction of solubility based on temperature was more precise and easier than prediction for newer compounds.

0.3 Weaknesses

One main issue for the paper was the distribution of the data itself, the solubility values that were recorded in standard unit $\frac{mol}{L}$ had to be converted to log scale to reduce the skewness of the distribution. The authors also said that there were minimal compounds in a log of solubility range between -7 to -5 which were compounds that had lower solubilities. Predictions of solubilities on such types of compounds would be difficult given the small sample size in the training set.

0.4 Relevance

This paper focused on the development of QSPR models which are similar to quantitative structure activity relationship (QSAR) models. Based on the description of QSAR models provided in the solubility assignment, it is interesting to note that the authors rule out the use of linear models and demonstrate the viability of machine learning methods over current empirical methods.The interdependence of different features on one another and the complexity that this brings to the relationship between structural properties and solubility makes it a little challenging to find appropriate data to develop models, but the utility of using molecular finger printing has been shown to be beneficial here.

The use of the ECFPs is also similar to the Tanimoto scores calculated in the similarity search assignment, and perhaps could be used in lead optimization when trying to design libraries to screen, after which lead optimization could involve development of such models to predict solubility of such candidates.