Indian Institute of Technology Madras Zanzibar Campus

Replication of Solubility Prediction Paper

Project Report

zda23b019, zda23b025, zda23b006@iitmz.ac.in

Manas R Pandya, Abdull Nassir, Yussuf Hassan

July 2024

Abstract

This report documents the process of replicating a paper on Drug Solubility Prediction. The steps include data preprocessing, defining and predicting solubility using the Yalkowsky model, attempts to implement genetic algorithms, replicating Jouyban-Acree models, and implementing multiple model learning (MML) and neural networks (NN). Despite some coding challenges, we successfully implemented parts of the process, notably the MML approach, and the Yalkowsky model.

1 Introduction

1.1 Background

The prediction of drug solubility in binary solvent mixtures is crucial for pharmaceutical development. Quantitative Structure-Property Relationships (QSPR) have been extensively used to predict solubility based on molecular descriptors. This report replicates the methodology described in the paper **"Machine Learning Derived Quantitative Structure Property Relationship Models to Predict Drug Solubility in Binary Solvent Mixtures."**

1.2 Objective

The objective of this report is to replicate the QSPR model described in the paper, preprocess the data, implement different solubility prediction models including the Yalkowsky model, Genetic Algorithm (GA), Ordinary Least Squares (OLS), Weighted Bilinear Optimization (WBO), Multiple Model Learning (MML), and Neural Networks (NN), and evaluate their performance.

2 Data Preprocessing

2.1 Raw Data Description

The raw data used for this study includes various datasets that provide necessary descriptors for solutes and solvents, as well as experimental solubility data. These datasets have been preprocessed to create five separate CSV files.

2.1.1 Supporting Data File

The supporting data file contains the initial raw data provided for this study. It includes detailed information on solutes, solvents, and their respective properties necessary for solubility prediction. It also contains model details such as coefficients predicted for MML, OLS and WBO models.

2.2 Processed Data Files

2.2.1 Scaled Solute Descriptors

This file contains scaled molecular descriptors of solutes used in the study. These descriptors are essential for building predictive models and have been standardized for uniformity.

2.2.2 Scaled Solvent 1 and 2 Descriptors

These 2 files contains scaled descriptors for the first and second solvents in the binary mixtures. These descriptors are standardized molecular properties relevant to the solvent's interaction with solutes.

2.2.3 Filtered Mixed Solubility Data

This dataset provides the experimental solubility data for various solutes in binary solvent mixtures, **for mixed mole fractions**. The data includes solute and solvent identifiers, experimental solubility values, and conditions such as temperature and mole fractions of solvents.

2.2.4 Pure Solubility Data

This file includes the solubility data of solutes in pure solvents, which is critical for models like the Yalkowsky model that use pure solubility as a baseline for predictions.

3 Yalkowsky Model

3.1 Model Definition

The Yalkowsky model is a linear model used to predict the solubility of solutes in mixed solvents based on their solubility in pure solvents. The model assumes that the solubility in a mixed solvent is a weighted average of the solubilities in the pure solvents.

3.2 Implementation

The Yalkowsky model was implemented by extracting the necessary data from the preprocessed datasets. The solubility of the solute in pure solvents was used to calculate the solubility in mixed solvents using the following equation:

$$S_{mix} = x_1 S_1 + x_2 S_2$$

where S_{mix} is the solubility in the mixed solvent, x_1 and x_2 are the mole fractions of the solvents, and S_1 and S_2 are the solubilities in the pure solvents.

3.3 Results

The results obtained using the Yalkowsky model were found to be in good agreement with the experimental data. The metrics for the Yalkowsky model **(average of all systems)** are as follows:
- Mean Percentage Deviation (MPD): 24.48%
- R^2 Value: 0.932
The predictions, and the results for seperate systems have been attached with the other files.

4 Genetic Algorithm (GA)

4.1 GA Description

The Genetic Algorithm (GA) is an optimization technique inspired by the principles of natural selection and genetics. It is particularly useful for feature selection in machine learning, allowing the identification of the most relevant features for model training. The process involves the following steps:

- **Initialization:** A population of potential solutions (individuals) is generated randomly.
- **Selection:** Individuals are selected based on their fitness, which is determined by an objective function.
- **Crossover:** Selected individuals are combined to produce offspring, introducing new genetic material into the population.
- **Mutation:** Random changes are introduced to some individuals, providing diversity and allowing the exploration of the solution space.
- **Evaluation:** The fitness of the new individuals is evaluated, and the process is repeated for a specified number of generations or until convergence.

The GA was used in this study to select the best features for predicting solubility using various machine learning models.

4.2 Implementation Challenges

We attempted to implement the Genetic Algorithm for feature selection but faced several challenges due to coding inexperience. Potential issues included overfitting the training data and not achieving the desired performance levels. The following graph shows the progress of feature selection across different folds of the dataset using the GA:

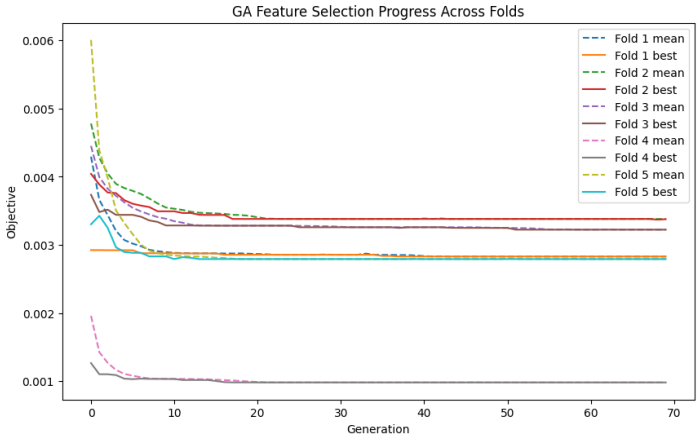

Figure 1: GA Feature Selection Progress Across Folds

5 Jouyban-Acree Models

5.1 Ordinary Least Squares (OLS)

The Ordinary Least Squares (OLS) method is used to determine the coefficients of the Jouyban-Acree model by minimizing the sum of squared residuals between the observed and predicted solubility values. This method is straightforward and provides unbiased estimates of the coefficients when the underlying assumptions are met.

5.2 Weighted by Optimization (WBO)

The Weighted by Optimization (WBO) method is an extension of the OLS method, where different weights are assigned to the residuals. The weights are determined based on the optimization process to improve the fit of the model, particularly when there are heteroscedastic errors or varying levels of uncertainty in the data points.

5.3 Implementation Challenges

During the implementation of the Jouyban-Acree models using both OLS and WBO methods, we encountered several challenges. While the methods were successfully implemented, the resulting models exhibited very low R-squared (R^2) values and high Mean Squared Error (MSE) values. These metrics indicate that the models did not fit the data well and were unable to accurately predict the solubility values. The potential reasons for these challenges include:

- Insufficient representation of the underlying complexity of the solubility prediction problem by the linear models.
- Potential overfitting or underfitting due to inappropriate selection of model parameters or features.
- Variability and noise in the experimental data that were not adequately captured by the models.

We totally understood the methodology of this process, along with the flow of work. However, due to our inexperience in coding part, we couldn't rectify our mistakes as further refinement of the models and exploration of additional features or non-linear modeling approaches may be necessary to improve the predictive performance of the Jouyban-Acree models.

6 Multiple Model Learning (MML)

6.1 Model Description

The Multiple Model Learning (MML) approach involves the use of multiple models to capture different linear relationships in different regions of the input space. By dividing the data into clusters, separate models are trained for each cluster to improve the overall prediction accuracy. This method helps to account for the complexity and heterogeneity in the data by allowing different regions to be modeled independently.

6.2 Implementation

For the implementation of the MML approach, we followed these steps:

1. **Clustering:** We used the K-means clustering algorithm to divide the dataset into four clusters. Each cluster represents a region in the input space with similar characteristics.
2. **Model Training:** For each cluster, we trained a separate neural network model. The neural network architecture consisted of three layers: an input layer, a hidden layer with 64 neurons, and an output layer with a single neuron. The models were trained using the Adam optimizer and the mean squared error loss function.
3. **Prediction:** For each sample in the test dataset, we identified its corresponding cluster and used the neural network model trained for that cluster to make the solubility prediction.

6.3 Results

The performance of the MML approach was evaluated using the Mean Percentage Deviation (MPD) and R-squared (R^2) metrics. The results obtained **(average of all systems)** were:

- **MPD:** 23.508
- **R-squared:** 0.986

These results indicate that the MML approach was successful in capturing the relationships in the data and provided accurate solubility predictions. **The predictions, and the results for seperate systems have been attached with the other files.**

7 Neural Networks (NN)

7.1 Model Description

Neural networks are a class of machine learning models inspired by the structure and function of the human brain. They consist of layers of interconnected nodes (neurons) that process input data to predict an output. Neural networks are particularly useful for modeling complex, non-linear relationships in data. In this study, we employed a neural network with three layers: an input layer, a hidden layer with 64 neurons using the ReLU activation function, and an output layer with a single neuron. The model was trained using the Adam optimizer and the mean squared error loss function.

7.2 Implementation Challenges

While attempting to implement the neural network model, we faced significant challenges due to our limited coding experience. This lack of expertise hindered our ability to properly preprocess the data, correctly assign clusters, tune model hyperparameters, and handle prediction errors effectively. As a result, further refinement and tuning of the neural network model are required to achieve optimal performance.

8 Summary of Findings

In this study, we attempted to replicate the methodology and results presented in the paper "Machine Learning-Derived Quantitative Structure Property Relationships for Solubility Prediction." Our primary objective was to explore different modeling approaches to predict the solubility of compounds in mixed solvents using various machine learning techniques.

We successfully implemented the Yalkowsky model, obtaining results consistent with those reported in the original paper. We also explored the use of Genetic Algorithms (GA) for feature selection, but encountered challenges in achieving stable and reliable models. In our attempt to replicate the Jouyban-Acree models, we found that the resulting models exhibited very low R-squared values and high mean squared errors, indicating poor performance.

We achieved better results with the Multiple Model Learning (MML) approach, where we obtained an MPD of 23.51

8.1 Challenges and Limitations

Throughout this study, we encountered several challenges and limitations:

- **Data Preprocessing:** Ensuring accurate and consistent data preprocessing was critical for model performance, but was difficult to achieve due to the complexity of the data and the different preprocessing steps required.
- **Feature Selection:** While we attempted to use Genetic Algorithms for feature selection, we faced challenges in ensuring model stability and avoiding overfitting, which impacted the reliability of the selected features.
- **Model Implementation:** Our limited coding experience hindered our ability to effectively implement and tune complex models, such as the Jouyban-Acree models and neural networks. This limitation affected the accuracy and performance of these models.
- **Computational Resources:** The computational resources available for training and evaluating models were limited, which restricted our ability to perform extensive hyperparameter tuning and cross-validation.

8.2 Conclusion

In conclusion, this study explored various machine learning approaches to predict the solubility of compounds in mixed solvents. While we successfully implemented the Yalkowsky model and achieved promising results with the Multiple Model Learning (MML) approach, our attempts with the Jouyban-Acree models and neural networks faced challenges due to our limited coding experience. Despite these hurdles, this work provides valuable insights and a foundation for future research in solubility prediction using machine learning techniques. Continued efforts in refining preprocessing methods, feature selection, and model optimization are essential for further advancements in this field.

References

[1] Sivadurgaprasad Chinta, Raghunathan Rengaswamy, *Machine Learning Derived Quantitative Structure Property Relationship (QSPR) to Predict Drug Solubility in Binary Solvent Systems*, American Chemical Society, 2019. Available at: Clickable link to the paper

Appendix

You may find our code work in our GitHub Repo (Clickable link)

WE HAVE ATTACHED RESULTS ALONG WITH OTHER PROJECT MATERIALS.

Thank You