Drug Solubility Predictions in Binary Solvents - An AI/ML perspective

IIT Madras, Zanzibar Campus

Chemistry Group Project

Nas-bah Khamis, Aradhya Goswami,
Aryan Rajput and Sai Adith Prakash

1. ABSTRACT :

This review analyzes a new QSPR method based on machine learning in the article for predicting the solubility of drugs in binary solvent mixtures. A genetic algorithm is used to analyze and obtain the final set of features. Single and multi-modelling strategies were used. Under the single model, Ordinary Least Squares and a Weight Based Optimization approach is used. Under the multi-model approach, solubility is assumed to be piecewise linear.

Quantitative Structure-Property Relationship (QSPR) is a mathematical model that correlates the physical properties of a molecule with its structural characteristics, represented by descriptors or features, which are numerical values reflecting various aspects of the molecule's structure, such as atom counts and topological surface area.

PROBLEM STATEMENT :

Accurately predicting drug solubility in binary solvent systems is a critical yet challenging task in pharmaceutical research. Achieving the desired drug concentration in circulation is crucial for eliciting the required pharmacological response. Since drugs must reach their receptors through aqueous media, aqueous-soluble drugs are favored for clinical applications. However, the increasing number of drugs failing commercialization due to poor aqueous solubility has highlighted the need for effective solubility enhancement strategies. Advanced predictive models and machine learning techniques have been employed to better understand and improve drug solubility.

The Jouyban-Acree model at it's initial stage was used to predict solute solubility when the temperature was constant. Another researcher called Hanaee regressed specific variables with a no intercept linear model (ie. intercept constrained to 0). This was further developed and the updated Jouyban-Acree model to predict solubility for varying temperatures as well.

The aim of this project is to map drug solubility in binary solvent systems to features like molecular weight, molar refractivity and so on, employing a modified version of the Jouyban-Acree model to do so. There are many ways to predict drug solubility in binary solvents, however in this research paper our focus is to find the most efficient and accurate model to do so , and to find it out we test various models to compare them and find the best one .

3. DATA DESCRIPTION :

The data we had was initially combined. We ran code to split the data to sort solubility data of mixed and pure compounds. We also manually exported files (using spreadsheets) for descriptors. A master file containing all features was used as well. Data was collected from the provided data set which included 63 binary systems of solutes and solvents at various temperatures.

The data included 27 different solutes and 21 solvents. The data had 12 descriptors namely with symbols :

AMR	Molar refractivity
Apol	Sum of the atomic polarizabilities
Si	Sum of first ionization potentials
McG_Vol	McGowan characteristic volume
VABC	van der Waals volume
MLFER_A	Overall solute hydrogen bond acidity
MLFER_BH	Overall solute hydrogen bond basicity
MLFER_S	Polarizability
MLFER_E	Excessive molar refraction
MW	Molecular weight
TopoPSA	Topological polar surface area
SolvAc SA	Sum of solvent accessible surface area

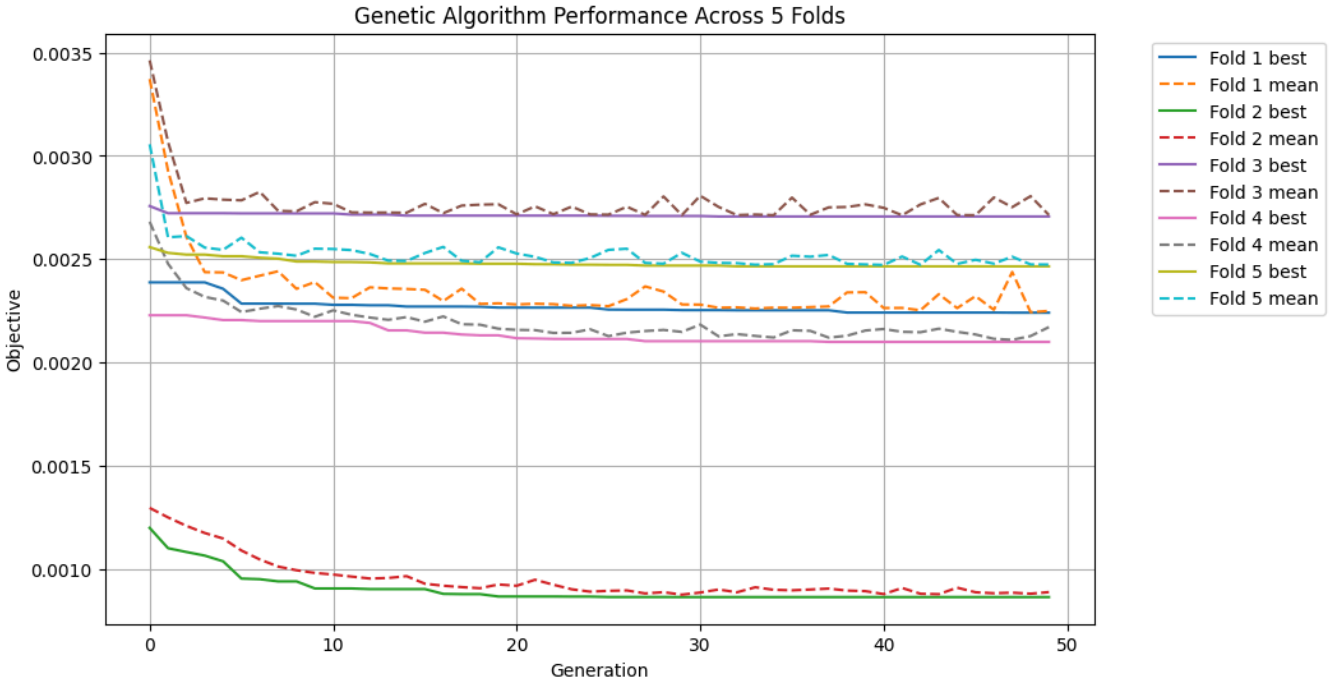
The descriptor values for the solutes , solvents and all the systems were also given.

4. EXPLORATORY DATA ANALYSIS :

Yalkowsky model accuracy statistics:

```
Average Metrics:
MAE: 0.0125077747798928201
RMSE: 0.0140100856316387 94
R2: -2.792461032597356
```

Genetic Algorithm's 5 folds.

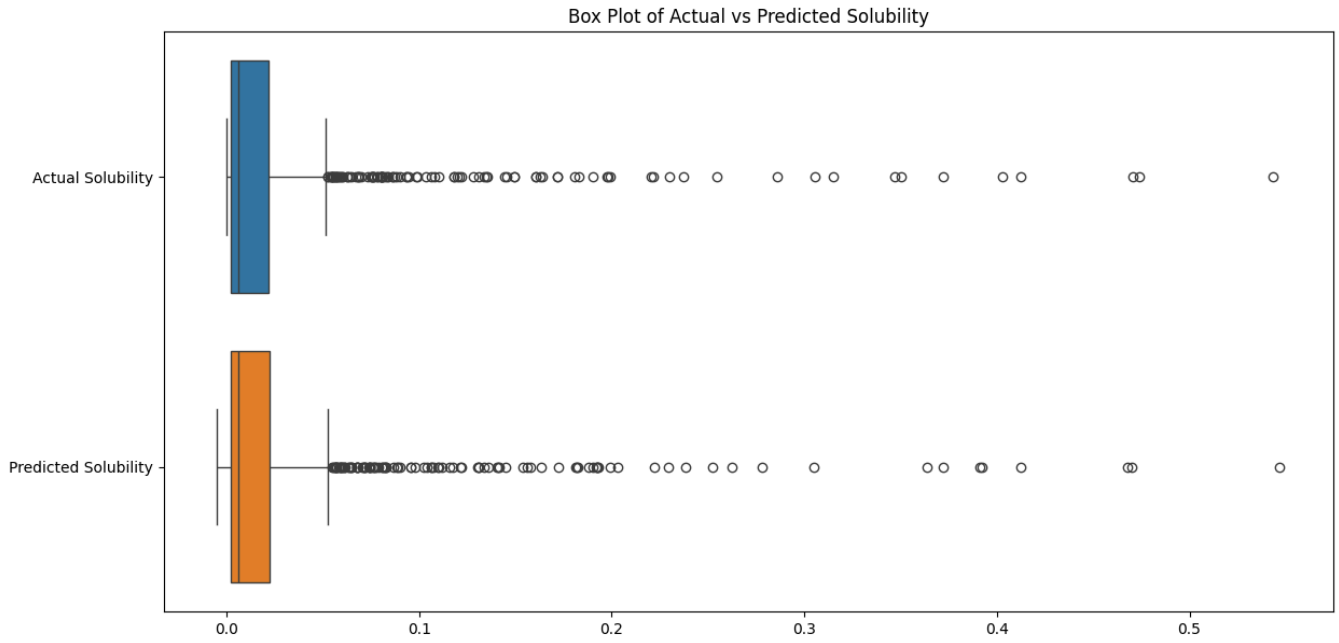


The third fold has seen considerable variance from the rest. It's dramatically lower than the rest of the folds. The parameters of the GA, such as mutation rate, crossover rate, and population size, might have a more pronounced effect on some folds than others - we chose to flag this fold.

OLS accuracy statistics after using GA:

```
Mean Absolute Error: 0.027660475449269948
Root Mean Squared Error: 0.039321947791755 23
R2: 0.1506659232637214
```

Piecewise Linear Function boxplots:



Accuracy Statistics of the piecewise linear function:

```
Average RMSE: 0.0013464765412472093
Minimum RMSE: 0.0010432813227424307
```

```
The R2 value is: 0.9795108694693381
The MAE value is: 0.0027439402969403497
```

5. METHODOLOGY:

The Yalkowsky Model:

Yalkowsky's model is a particular case of the General Solubility Equation whereby, similar to a weighted average, the respective solubilities and mole fractions are multiplied for 'x1' and 'x2' and then added together. This was a simple mathematical model and whilst it definitely did show some similarities with the actual data, we wanted to use a better model with better accuracy.

Single Linear Model + Genetic Algorithm:
We decided to use a linear model (Ordinary Least Squares) to predict experimental solubility. But instead of using all features, we used a Genetic Algorithm to filter the ones most correlated with solubility. In short, this algorithm works by selecting data samples and 'crossing them' to form a daughter solution which can then be 'mutated' - or modified. This process will identify the better features, of which we fed into the OLS model. But not only did this not take considerably more time, the R2 was constantly low and was varying in the 0.1-0.3 range - considering R2 is out of 1, the variance of this model is colossal.

Multiple model approximation : To experiment within the **multiple model space**, we decided to make a piecewise linear model by using the prediction error based fuzzy clustering approach.
This Algorithm is used to identify the various linear models hidden in a single linear model approximation. This algorithm divides m data samples in n different linear models . The n models are made by making clusters of data and minimizing the objective of the algorithm. By using gradient of the function we can move in the direction of minimization and identify the cluster center. After identifying the cluster center we can take any other point k and put it in the desired cluster by using the cosine metric between the clusters i and k

Then KNN testing strategy is used to predict the drug solubility. First, the Euclidian distance from the test sample to

all samples in the training data are evaluated. Then the nearest K samples to test sample and their corresponding models are identified.

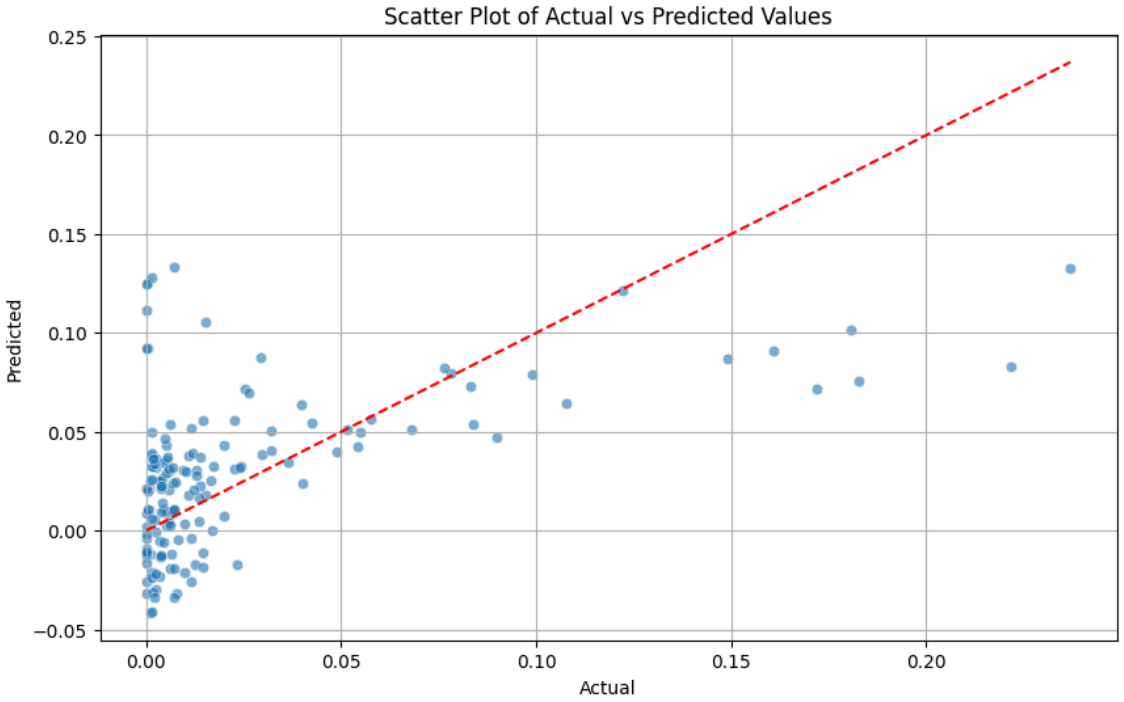
By identifying clusters that exhibit distinct patterns or relationships between variables, our predictions were much more accurate

Results:

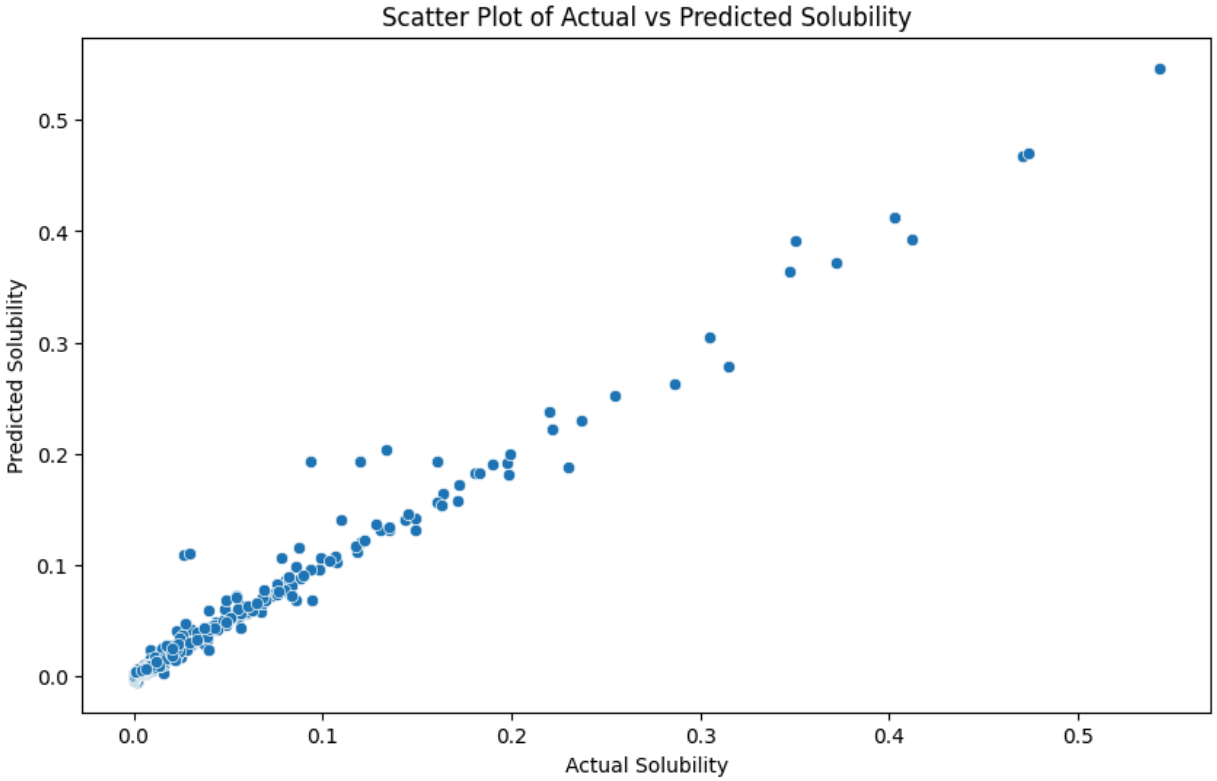
	Yalkowsky's GSE	OLS + GA	Piecewise Linear
MAE	**0.013**	**0.028**	**0.0027**
RMSE	**0.014**	**0.039**	**0.0013**
R2	**-2.79**	**0.151**	**0.98**

Yalkowsky's GSE's scatter plot

Genetic Algorithm and Ordinary Least Squares Scatter Plot



Linear Piecewise Scatter Plot: Greater accuracy shown



Challenges Faced:

We had a number of challenges whilst working on this research paper. Primarily, the running of the Genetic Algorithm to select the best features was time intensive. So was the case with the piecewise linear model, as the iterations took a lot of computational power. With online notebooks, only a limited amount of GPU power is available and therefore computationally it was intensive and took a lot of time. This means that if the code contains errors, it will take longer to find out. Furthermore, implementing the fuzzy based clustering algorithm was a challenge in which we got help from Mr. Madhusudhan Verma from our MTech batch.

In conclusion, both the single models fare poorly against the piecewise linear model approach. With an R2 value of around 0.98 the piecewise linear model was by far the best model among all . Although we did try to use support vector regression, our inexperience proved to be a strong disadvantage as we weren't able to complete it on time. The piecewise function has a plethora of applications in the pharmaceutical industry. For us, we believe the main advantage would be that this more efficient and cost effective which will allow more pharma firms into the business, increasing competition. With this, pricing will also decrease, which will benefit everyone - especially those who cannot afford it if medicines were expensive.