How well machine learning models can predict the aqueous solubility of small organic molecules?

Vasileios A. Tatsis

May 25, 2018

Abstract

Aqueous solubility is considered to be one of the important properties in drug discovery. In this work, the performance of several machine learning methods in inferring the aqueous solubility of a small organic molecules' is evaluated.

Domain Background

Solubility is the property of a solid, liquid, or gaseous chemical substance called solute to dissolve in a solid, liquid, or gaseous solvent to form a homogeneous solution of the solute in the solvent. The solubility of a substance fundamentally depends on the solvent used as well as on temperature and pressure. The extent of solubility of a substance in a specific solvent is measured as the saturation concentration where adding more solute does not increase its concentration in the solution.

Aqueous solubility prediction is important in drug discovery. Early identification of a compound with poor aqueous solubility in a drug discovery cascade can minimize the risk of failure. Currently, I am working in the drug discovery field and I am very interested in using predictive models for the design or optimization of small organic drug molecules.

Problem Statement

The aqueous solubility of a drug is a significant factor for its bioavailability. Many drugs are administered via the oral route, their absorption and metabolism in organisms are closely related to its aqueous solubility.

A model capable of accurately predicting the aqueous solubility of small organic molecules will be very useful in the Drug Discovery field. In this work, I will try to compare different machine learning models in predicting the aqueous solubility of small organic molecules. As a sole descriptor I will use the compounds' structural fingerprints. Other helpful descriptors could be used, like physicochemical properties or geometric features.

Furthermore, RDKit (an open source cheminformatics Python library) will be used to manage the chemical information in this project. Four standard metrics: i) Pearson correlation coefficient (R), ii) mean absolute error (MAE), iii) root mean square error (RMSE) and iv) coefficient of determination (R^2) will be used to compare the performance of the models.

Data sets and Inputs

For the training and testing of the machine learning models I will employ the Huuskonen data set (ref. 1). This data set contains 1026 organic molecules by Jarmo Huuskonen (ref. 2). Molecules are listed in smiles format along with their aqueous solubility values, expressed in log mol/L at 20-25 degrees, and their octanol-water partition coefficient values (logP).

I will use as input to the machine learning models only the structural information (2D graph) from this data set to infer the aqueous solubility of the compounds contained in the data set. *The simplified molecular-input line-entry system (SMILES) is a specification in form of a line notation for describing the structure of chemical species using short ASCII strings* (Wikipedia). For example, cyclopropene is usually written $C1 = CC1$.

For this purpose the chemical structure information will be converted into feature vectors of fixed length (fingerprints). Before training the machine learning models, the available fingerprints in RDKit will be tested in order to select the one that can represent chemical structure with a minimum loss of information. Molecular fingerprints encode molecular structure in a series of binary digits (bits) that represent the presence or absence of particular substructures in the molecule. But, fingerprints do not include full structural data (such as coordinates). The cross-validation technique will be also employed.

RDKit, an open-source cheminformatics Python library, will be used to drive the conversion of the chemical structure into feature vectors of fixed length (structural fingerprints).

Solution Statement

A public data set that contains 1026 organic molecules and their aqueous solubility values, expressed in log mol/L at 20-25 degrees, will be used in this study. Structures in this data set are given in SMILES format, strings that encode molecular graphs into a string of characters. RDkit, an open source cheminformatics Python library, will be used to convert the SMILES to feature vectors of fixed length (structural fingerprints). Fingerprints will be used as the sole descriptor and input in the machine learning models that will be employed. The target value will be the aqueous solubility values of these compounds.

Benchmark Model

The outcome of this study will be compared to the experimental results and to the other corresponding results found in literature (ref. 4). Frohlich et al. (ref 4), using the same data set, report a squared correlation coefficient of 0.90 for an 8-fold cross-validation, using support vector machines with a radial basis function kernel.

Evaluation Metrics

In order to compare the performance of the machine learning models used in this work, I will use three standard metrics: i) Pearson correlation coefficient (R), ii) mean absolute error (MAE), iii) root mean square error (RMSE), and iv) coefficient of determination (R^2) defined by:

$$R = \frac{\sum_{i=1}^{n}(t_i - \overline{t})(p_i - \overline{p})}{\sqrt{\sum_{i=1}^{n}(t_i - \overline{t})^2}\sqrt{\sum_{i=1}^{n}(p_i - \overline{p})^2}}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|t_i - p_i| \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{i}(t_i - pi)^2}$$

where t_i is the target value (experimentally measured) and p_i is the predicted value for a compound (data point) i.

Project Design

The aim of this project is to evaluate the performance of several machine learning models in predicting the aqueous solubility of small organic molecules. Aqueous solubility is a critical property since it is related to the absorption and metabolism of oral drugs in human body. In this study, a public data set containing the 2D structures (in SMILES format) of small organic molecules and their aqueous solubility values will be used as input. In this work, only one dimension or feature (the structural fingerprints) will be used as input to the machine learning models for each data point. An analysis of the data set will be the first step in this study. Grid searches to tune the hyperparameters of the machine learning models like RandomForestRegressor, KernelRidge, and GaussianProcessRegressor will be applied. Simple learning models like PLS will be employed.

Also, XGBoost and LightGBM models will be evaluated in this work. I will explore how chemical diverse is the data set and highlight possible implications (if any) of the chemical diversity in the predictability of the machine learning models.

References

1. http://cheminformatics.org/data sets/
2. J. Chem. Inf. Model., 2013, 53 (7), 1563–1575
3. SAR and QSAR in Environmental Research, 2008, 19 (3-4), 191-212
4. QSAR Comb. Sci., 2004, 23, 311-218