AI in Chemical Engineering Report: Mol2vec Embeddings for Lipophilicity Prediction

Fiammetta Caccavale
fiacac@kt.dtu.dk

1 Introduction

This work aims to investigate the lipophilicity (logP) property of compounds. The molecules are embedded in high dimensional space through Mol2vec [1], an unsupervised machine learning approach to learn vector representations. The vector representations of the molecular structures are then used to train various machine learning models on a regression task, specifically predicting the lipophilicity of compounds.

Compound representation through Mol2vec

A meaningful knowledge representation is one of the keys to predictive models. The methods used to encode molecular structures have long been researched and optimized throughout the years, and a variety of descriptors and molecular fingerprints (FP) have been developed. Recent development in various fields of machine learning, such a natural language processing (NLP), have paved the way for the development of models such as Mol2vec, an NLP-inspired technique that build on Word2vec [2]. Mol2vec is an unsupervised method that learns compound structures derived from the Morgan algorithm as "words" and compounds as "sentences". The high-dimensional vectors of substructures obtained through Word2vec are the summed to obtain compound embeddings [1]. The obtained vectors of molecular structures can be used to train machine learning algorithms, for similarities search and clustering.

Property prediction

Estimating molecular properties under normal laboratory conditions requires large experimentation efforts, time and a high budget, since most molecules can be difficult to obtain and expensive to synthesize, leading to an increasing necessity for data-drive approaches for property estimation. Data-driven approaches for property prediction build on underlying relationships between molecular features and their physical properties. This work focuses on lipophilicity prediction. Lipophilicity (logP) is the ability of a chemical compound to dissolve in lipids, oils and non-polar solvents. This property is usually evaluated through a distribution coefficient P, which is the ratio at equilibrium of the concentration of a compound between two phases, an oil and a liquid phase [3]. The greater the logP value, the greater is the lipophilicity.

2 Data

The dataset provided contains 12193 compound structures represented as SMILES (Simplified Molecular Input Line Entry System) [4]. The dataset is pre-processed to exclude compounds duplicates. The original dataset presents 179 duplicated SMILES, therefore it was decided to keep the unique compounds and the average value of the duplicated logP property. After pre-processing, the vector representations of the molecular structures, extracted with RDkit[1], were extracted through Mol2vec, resulting in a corpus of shape [12005, 300], where

[1] RDkit: https://www.rdkit.org/ [Accessed on 10/06/2022].

the first dimension is the number of examples and the second the embedding dimension. For visualization purposes and to better inspect the data in lower-dimensional space, various dimensionality reduction methods were applied, such as in Figure 1[2].

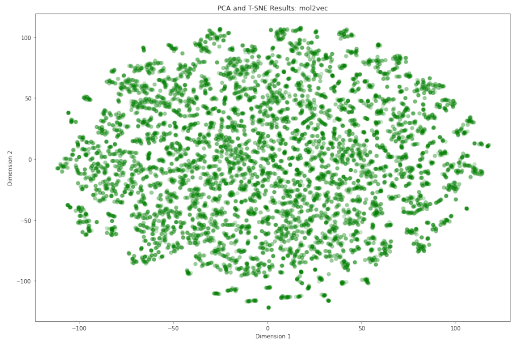

Figure 1: Dimensionality reduction of mol2vec embedding using PCA (to 30 components) and T-SNE (to 2 components).

To train the models, the dataset was split into three sets, training, validation and test, with 80/10/10 split respectively. Hyper-parameter tuning is done on the validation set through manual grid search of selected parameters. The results reported were calculated on an unseen (hold-out) test set, to prevent model bias.

3 Methods

This work proposes an empirical comparison of different machine learning algorithms evaluated using Mol2vec as compound features for lipophilicity prediction. The models evaluated are linear regression (LinR) which is used as baseline to benchmark the other methods, support vector machine for regression (SVR) with radial basis function kernel, extreme gradient boosting algorithm (XGBoost), ensemble of the previous three models (Ensemble), and neural networks such as a multilayer perceptron (MLP), a three layers deep neural network (DNN) with rectified linear unit activation, and bidirectional long-short term memory neural network (LSTM). The linear regression and SVR models were trained using the scikit-learn implementation[3] with default parameters. The XGBoost[4] presents the following hyper-parameters: *n_estimators=1000, max_depth=6, eta=0.3, subsample=0.7, colsample_bytree=0.8, reg='squarederror'*. The DNN is trained for 300 epochs, the LSTM was trained for 500 epochs. The algorithms are evaluated through various metrics, such as mean absolute error, mean squared error, root mean squared error and coefficient of determination (R^2).

4 Results and Discussion

All the implemented models present a high R^2 score on the logP property prediction, as shown in Table 1. These results suggest that the molecule vectors obtained through Mol2vec are a good representation of the compounds and manage to keep the *semantical* information of the substructures and to successfully learn underlying relationships between molecular features and their physical properties. When comparing the machine learning to the deep learning approaches, the latter achieve higher performances than simple machine learning models across all evaluation metrics. This is probably due to the more complex architecture of deep learning approaches that leverages on the high-dimensionality of the input. The lower performance of the bidirectional LSTM compared to the other neural networks is surprising, however more hyper-parameter tuning would probably be beneficial for this model and increase its performance.

[2]The compounds and their respective vectors are saved in two .tsv files. For further details, refer to the .ipynb notebook

[3]Sklearn: https://scikit-learn.org/stable/ [Accessed on 10/06/2022]

[4]XGBoost: https://xgboost.readthedocs.io/en/stable/ [Accessed on 10/06/2022]

	LinR	SVR	XGBoost	Ensemble	MLP	DNN	LSTM
MAE	0.479	0.443	0.435	0.409	0.353	0.330	0.365
MSE	0.393	0.377	0.365	0.301	0.231	0.221	0.252
RMSE	0.627	0.614	0.604	0.549	0.481	0.470	0.502
R2	0.888	0.892	0.896	0.914	0.934	0.937	0.928

Table 1: Comparison of selected models on lipophilicity (logP) prediction on hold-out set.

Figure 2: DNN model performance: observations vs predictions.

Limitations

Due to time constraints and computational efficiency, the current results are obtained over one run. However, future work will prevent model selection bias and ensure results stability by performing cross-validation. The LSTM did not manage to converge, although the model would probably manage to achieve the highest accuracy with more hyper-parameter tuning.

5 Conclusion

This work proposes and empirical comparison of different machine learning algorithms evaluated using Mol2vec as compound features for lipophilicity prediction. Both simple algorithms such as linear regression, used as baseline, than more complex neural networks as LSTM achieve high performance on the various evaluation metrics. The deep learning approaches leverage on the high-dimensionality of the input and reduce the error further. Overall, this work suggests that the Mol2vec model manages to learn meaningful representations of the molecules and their underlying properties.

References

[1] S. Jaeger, S. Fulle, and S. Turk, "Mol2vec: Unsupervised machine learning approach with chemical intuition," *Journal of Chemical Information and Modeling*, vol. 58, no. 1, pp. 27–35, 2018. PMID: 29268609.

[2] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[3] B. Chandrasekaran, S. N. Abed, O. Al-Attraqchi, K. Kuche, and R. K. Tekade, "Chapter 21 - computer-aided prediction of pharmacokinetic (admet) properties," in *Dosage Form Design Parameters* (R. K. Tekade, ed.), Advances in Pharmaceutical Product Development and Research, pp. 731–755, Academic Press, 2018.

[4] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 1, pp. 31–36, 1988.