Final Seminar Report

Group 2, Dec 11

Ming Jiang | mingj@kth.se

Sihan Chen | sihanc@kth.se

Gengcong Yan | gengcong@kth.se

Table of Contents

01 Data Preparation

02 Feature Sets Selection

03 Hyper-parameter Optimization

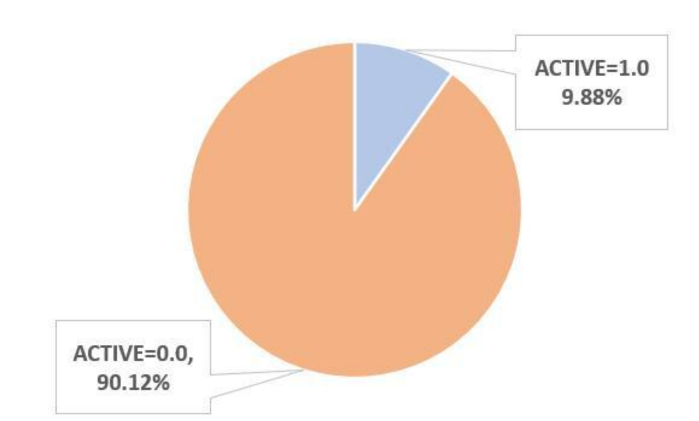
04 Future Work

01

Data Preparation

Raw Data Sample

Label Distribution



Imbalanced Data

*Only around **10%** compounds' labels are "ACTIVE".

*Influences:
(1) The sampling for Trees Generation -> Some parameter's tuning
(2) Evaluation index -> AUC instead of Accuracy

Data Separation

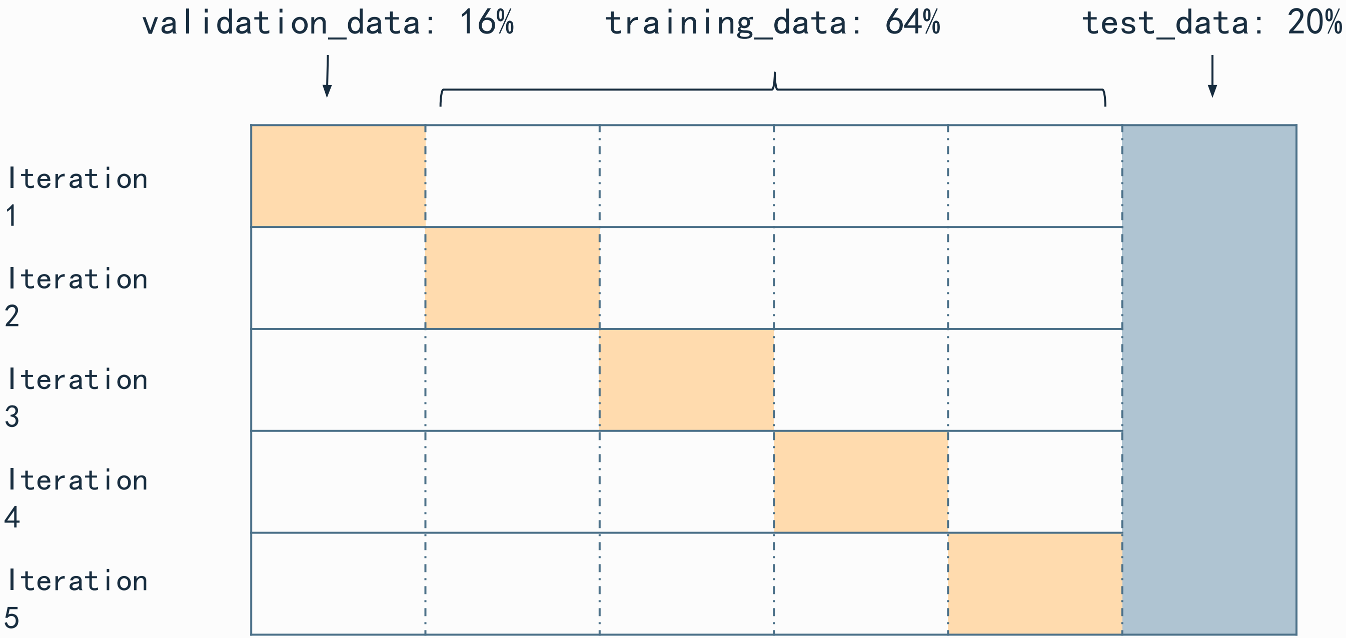


Fig. 1 k-folds(k=5) Cross-validation on Dataset

02

Feature Sets Selection

Basic Features of Chemical Compounds

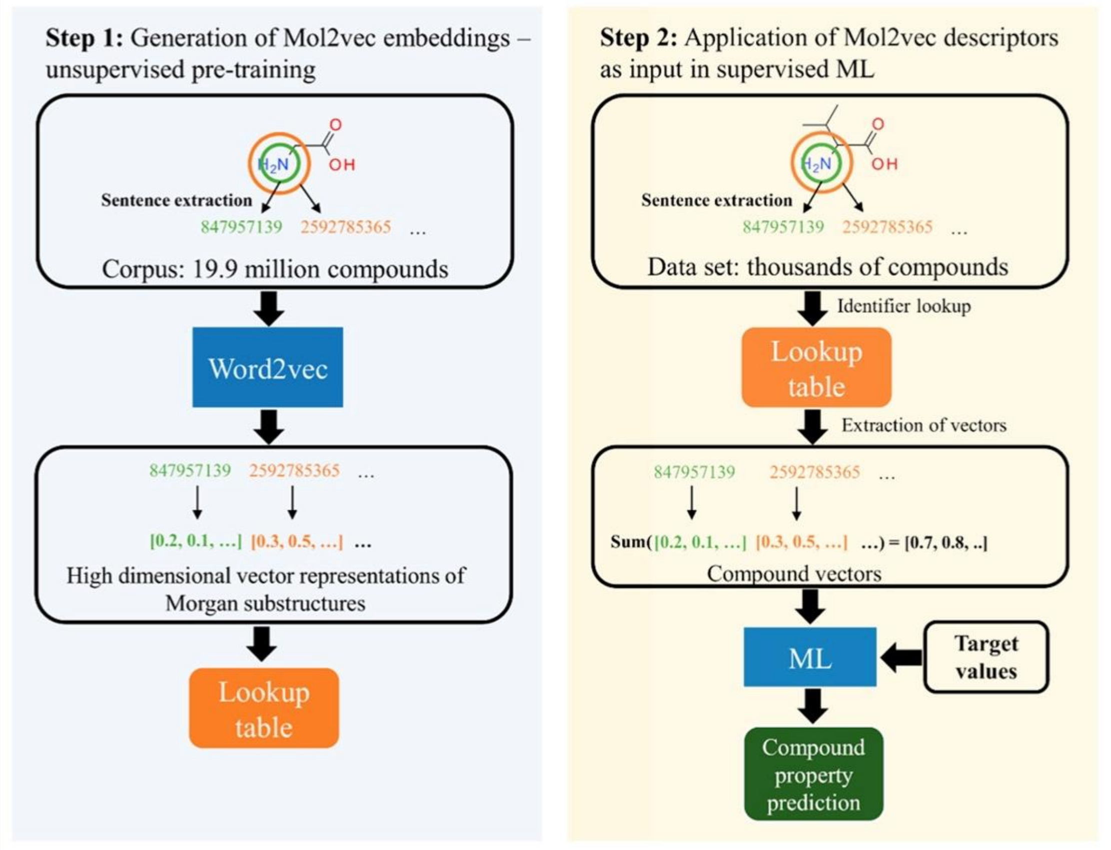
- Number of atom
- Molecular weight
- Number of N functional groups attached to aromat
- Number of halogens
- Number of aliphatic rings for a molecule
- Number of aromatic rings for a molecule

Three Kinds of Fingerprints

	Description	Parameters	Shape
ECFP*	Based on atom properties	(radius=2, nbits=124)	(, 124), ndarray
FCFP*	Based on pharmacophoric properties	(radius=2, nbits=124, useFeatures=**True**)	(, 124) , ndarray
mol2Vec	Based on word2vec		(, 300), ndarray

*Extended Connectivity Fingerprints
*Functional-Class Fingerprints

Mol2vec*


Step 1: Generation of Mol2vec embeddings – unsupervised pre-training
O
H_2N
OH
Sentence extraction
847957139 2592785365 ...
Corpus: 19.9 million compounds
Word2vec
847957139 2592785365 ...
[0.2, 0.1, …] [0.3, 0.5, …] ...
High dimensional vector representations of Morgan substructures
Lookup table
Step 2: Application of Mol2vec descriptors as input in supervised ML
O
H_2N
OH
Sentence extraction
847957139 2592785365 ...
Data set: thousands of compounds
Identifier lookup
Lookup table
Extraction of vectors
847957139 2592785365 ...
Sum([0.2, 0.1, …] [0.3, 0.5, …] …) = [0.7, 0.8, ..]
Compound vectors
ML
Target values
Compound property prediction

*Mol2vec is an unsupervised machine learning approach to learn vector representations of molecular substructures.

Baseline

	Random Forest	lightGBM	XGBM
Basic	0.587	0.640	0.627
FCFP	0.798	0.741	0.682
ECFP	0.784	0.740	0.687
m2v	0.761	0.764	0.726
Basic + FCFP	0.799	0.753	0.694
Basic + ECFP	0.796	0.744	0.696
Basic + m2v	0.773	0.773	0.732
Basic + FCFP + ECFP	**0.812**	0.772	0.715
ALL	0.786	0.784	0.741

Feature Selection

Model: Random Forest

Feature Set: { *basic+ECFP+FCFP* }

Feature Importance:

* We ran the PCA and only **one** column of feature was dropped.

03

Hyper-parameter Optimization

Hyper-tuning in Random Forest

Model: Random Forest

Feature Set: { *basic+ECFP+FCFP* }

```
{'class_weight': ['balanced_subsample']*,
 'criterion': ['entropy'],
'max_depth': [30, 40],
'n_estimators': [225, 250, 300, 350, 400],
'oob_score': [True]}
```

- Grid Search for the best parameters combination.
- 8 groups of parameters with 5-folds cross-validation in total.

*This parameter can help deal with the ***imbalanced data.***

Hyper-tuning in Random Forest

Table： The performance（AUC） of Random Forest on 5-folds Cross-validation with different parameters

max_depth	n_estimators	split0_test_score	split1_test_score	split2_test_score	split3_test_score	split4_test_score	mean_test_score	std_test_score
30	225	0.8186	0.8071	0.8139	0.8136	0.8182	0.8143	0.00415681
30	300	0.8218	0.8137	0.8160	0.8180	0.8229	0.8185	0.00346156
30	**400**	**0.8259**	**0.8147**	**0.8171**	**0.8211**	**0.8216**	**0.8201**	**0.00385479**
40	225	0.8196	0.8098	0.8150	0.8164	0.8190	0.8160	0.00352462
40	300	0.8255	0.8100	0.8159	0.8176	0.8194	0.8177	0.00502906
40	400	0.8269	0.8100	0.8172	0.8192	0.8238	0.8194	0.00581308
30	250	0.824	0.810	0.814	0.819	0.821	0.818	0.00480932
30	350	0.827	0.812	0.819	0.824	0.824	0.821	0.0051482
40	350	0.827	0.814	0.819	0.819	0.824	0.820	0.00473505
40	250	0.822	0.807	0.817	0.816	0.820	0.816	0.00521445

Model Comparison

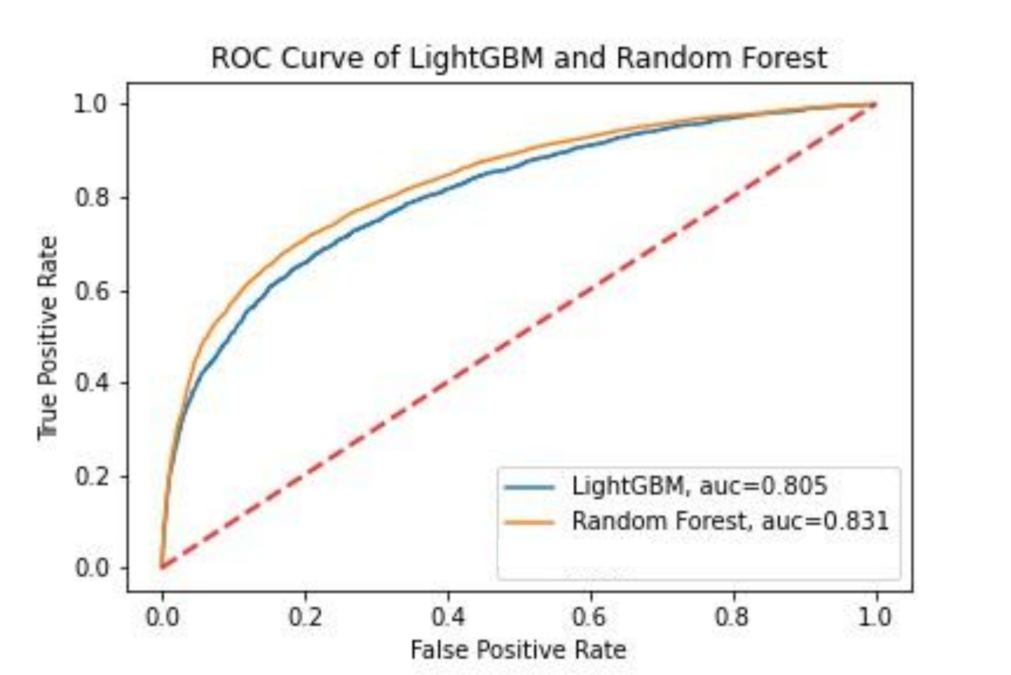


Figure: The performance(AUC) comparison of Random Forest and lightGBM with their relative best feature sets

Hyper-tuning in Random Forest

```
{'class_weight':  'balanced_subsample',
 'criterion':  'entropy']
'max_depth':  30,
'n_estimators':  400]
'oob_score':  True}
```

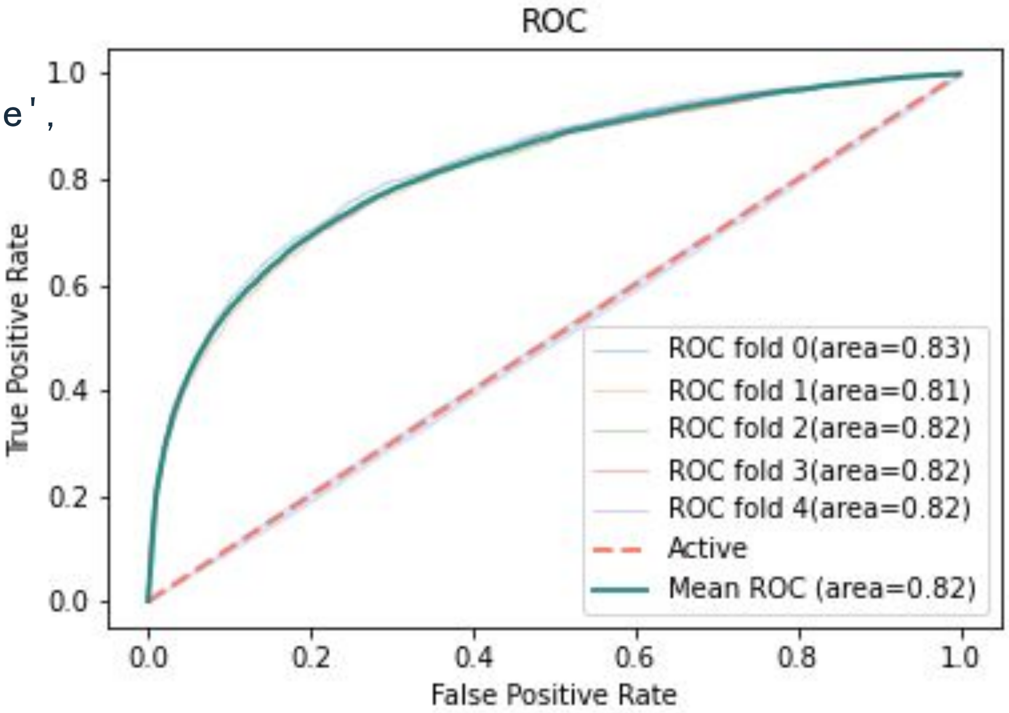


Figure: The performance(AUC) of Random Forest on 5-folds Cross-validation with the final parameters

04

Future Work

Future Work

- Consider to combine the results outputted by multiple models.

- Consider the possible prediction based on the data structure of chemical compounds.

References

[1] “Rdkit,” https://www.rdkit.org/, accessed December 9, 2020.

[2] D. E. Ratnawati, Marjono, and S. Anam, “Prediction of active compounds from smiles codes using backpropagation algorithm,” in AIP Conference Proceedings, vol. 2021, no. 1. AIP Publishing LLC, 2018, p. 060009.

[3] “Computing extended connectivity fingerprints,” https://depth-first.com/articles/2019/01/11/ extended-connectivity-fingerprints/, accessed December 9, 2020.

[4] 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.

[5] “Open source ecfp/fcfp circular fingerprints in cdk,” https://cheminf20.org/2014/02/21/ open-source-ecfpfcfp-circular-fingerprints-in-cdk/, accessed December 9, 2020.

[6] D. Weininger, A. Weininger, and J. L. Weininger, “Smiles. 2. algorithm for generation of unique smiles notation,” Journal of chemical information and computer sciences , vol. 29, no. 2, pp. 97–101, 1989.

Q & A

Thank you！