Project Report

Team member:

Yuji

Kaidi

Rui

CONTENT

CONTENT

01

Introduction

1.1 SMILES: Simplified Molecular Input Line Entry Specification

It is a real language structure, uses vocabulary (atomic and bond symbols) and grammatical rules to describe the structure of chemical compounds.

It is a popular tool applied in Machine Learning and Deep Learning tasks.

This Project:

In this project, we develop a predictive model by generating and selecting features from given training set and making optimization among different learning algorithms, with maximized **AUC.**

Introduction

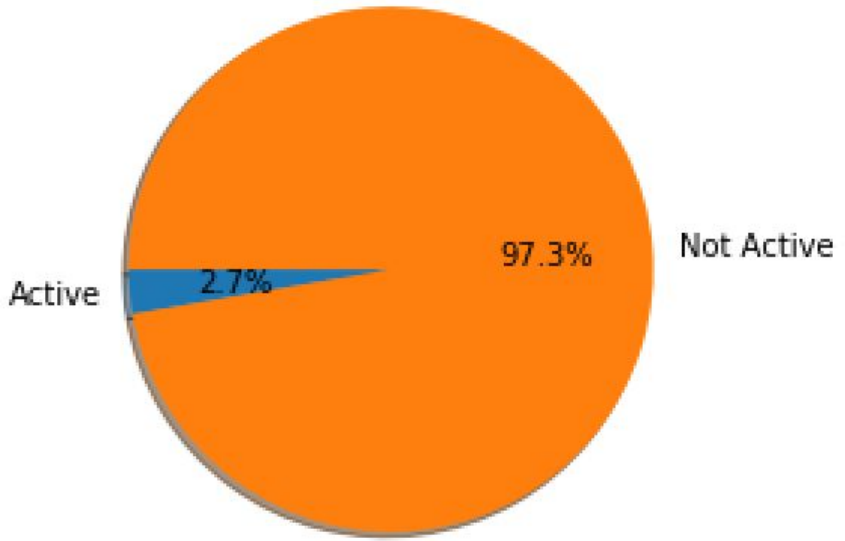
1.2

Data

Train Set Size: 19125 | Test Set Size:6375

SMILES	ACTIVE
Cl.O=C1C(CN(CC2:C:C:C:C:C:2)CC2:C:C:C:C:C:2)CCCC1CN(CC1:C:C:C:C:C:1)CC1:C:C:C:C:C:1	0
CCOC(=O)C1OC2:C:C:C([N+](=O)[O-]):C:C:2C1(C)O	0
COC1:C:C:C:C2:C:1OC1(N3CCOCC3)CCCCC1C2C[N+](=O)[O-]	0
CC1:C:C:C(S(=O)(=O)NC(=O)NCCSSCCNC(=O)NS(=O)(=O)C2:C:C:C(C):C:C:2):C:C:1	0
COC1:C:C:C:C:C:1NC(=S)NC=C1C(=O)NC(=O)NC1=O	0
C[N+]12CCC34C5:C:C:C:C:C:5NC3C3C(CC(=O)O)OCC=C(C1)C3CC42	0
O=C(O)C(NCC1:C:C:C:C:C:1)C(NCC1:C:C:C:C:C:1)C(=O)O	0
CC(C)=NNC=NC(C#N)=C(N)C#N	0
COC1:C:C:C(CC2C(=O)OCC2CC2:C:C:C(OC):C(OC):C:2):C:C:1	0
CCOCNC=O	0
CS(=O)(=O)OCCC(=O)N1CCN(C(=O)CCOS(C)(=O)=O)CC1	0
CC(CC(O)(C(F)(F)F)C(F)(F)F)=NNC(N)=S	0
CC(NC1=NCCO1)C1:C:C:C:C([N+](=O)[O-]):C:1	0
COC1:C:C(CC(C)NC(C)C2:C:C:C:C:C:2):C:C(OC):C:1	0
COC1:C:C:C(C=CC(=O)C2:C:C:C(Br):C:C:2):C:C:1	0
CN(C1:C:C:C:C:C:1Cl)S(=O)(=O)C1:C:C:C:C:C:1	0
CC1(C)OC(=C(C#N)C#N)C(C#N)=C1C=CC1:C:C:N:C:C:1	0
N#CC(N)=C(C#N)NC(=O)C(=O)NC(C#N)=C(N)C#N	1
COC1:C:C:C(P2(=S)OC3:C:C:C(C(C4:C:C:C:C:C:4)(C4:C:C:C:C:C:4)C4:C:C:C:C:C:4):C:C:302):C:C:1	0

Data Sample



Imbalanced Data

02

Modeling Method

2.1

Features Selection	Details
Morgan FingerPrint	**AllChem.GetMorganFingerprintAsBitVect(x,2,nBits=512)**
Molecular Descriptors	**Chem.MolFromSmiles():** a SMILES string and it will be assigned an object representing the corresponding chemical compound, for which various properties may be derived. **GetNumAtoms():** the number of Atoms without Hydrogen **CalcExactMolWt():** the molecule's exact molecular weight **fr_Al_COO():** the number of aliphatic carboxylic acids **Chem.AddHs().GetNumAtoms():** the number of Atoms with Hydrogen **BondType():** type of the bond as a BondType
Mol2vec	vector representations of molecular substructures

2.1 More Detail of Mol2vec

Inspired by natural language processing techniques such as Word2vec model, Mol2vec proposed by Sabrina Jaeger et al. (2018) is an unsupervised machine learning approach to learn vector representations of molecular substructures. By summing up vectors of the individual substructures, compounds can finally be encoded as vectors and, for example, feed into supervised machine learning approaches to predict compound properties

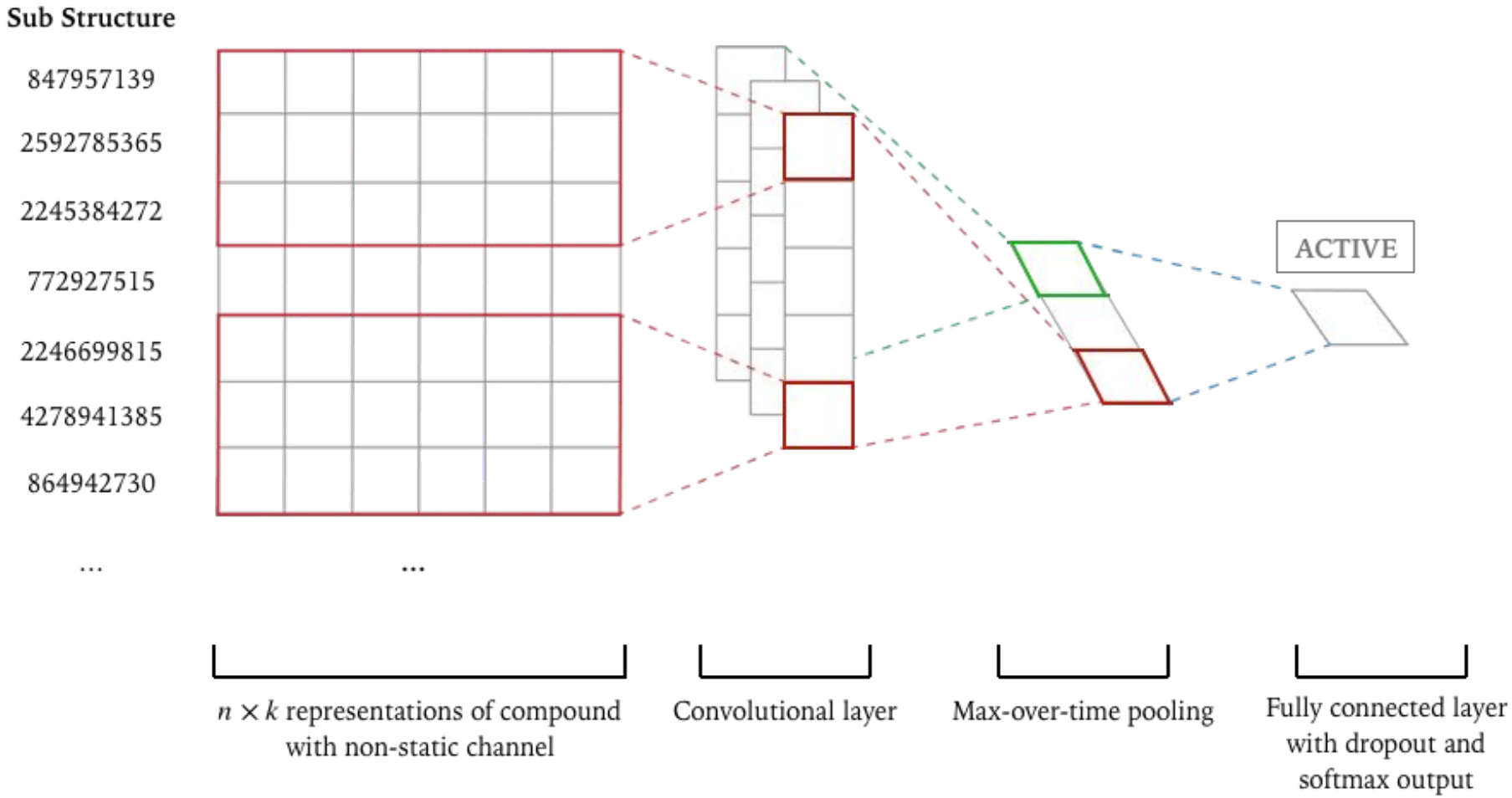
2.2

Training data – Validation data –Test data

Data type	percentage	number
All data	100%	19125
Training data	64%	12240
Validation data	16%	3060
Test data	20%	3825

2.3

Mol2vec + TextCNN + Fine Tune



03

Result & Analysis

3.1

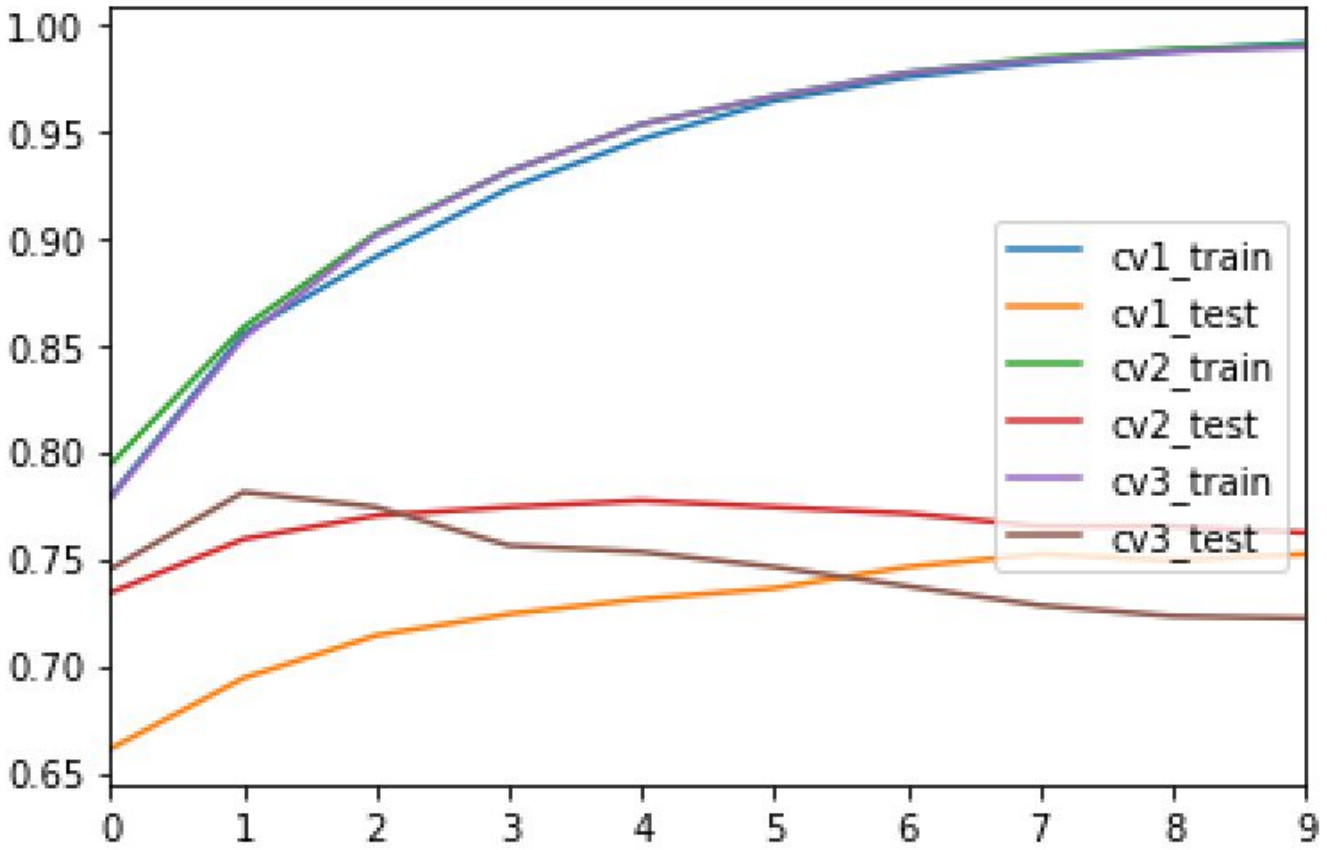
Models

	Logistic Regression	Random Forest	lightGBM
Morgan FingerPrint	0.7046286151365431	0.7681984268083434	0.7523230934612284
Molecular Descriptors	0.6726053273520353	0.7102683831889691	0.7240648192173319
Mol2vec	0.7032771380728919	0.7537314204204825	**0.8103125710623694**
All	0.7077863684286688	0.760380274119860 2	0.8073176151985613

Considering about the AUC score on 5 folds CV and overfitting, we choose lightGBM as our model and Mol2vec as our best features

3.2

Mol2vec + TextCNN + Fine Tune



Average AUC: 0.77

04

Further Study

4.1

In further study, we could

1. Tune parameters better
2. Try to apply some other effective models such as XGBoost and AdaBoost.
3. Apply Voting algorithm to combine weighted different models to improve the prediction.
4. Use combined two kinds of features to fit models(as in this project we just take the single kind of feature and all features into account)
5. Image classification on compounds' images

Thanks for watching