Drug Analysis by Using Machine Learning And Deep Learning

Xueli Zhou
Columbian College of Arts and Sciences
The George Washington University
Washington D.C.
Email: xueli81993@gmail.com

Siyuan Hu
Columbian College of Arts and Sciences
The George Washington University
Email: ....com

Ruoyu Wang
Columbian College of Arts and Sciences
The George Washington University
Email: ....com

***Abstract*—This project aims to find some useful informations from medical data. At first, we give an introduction about this project, and talk about some related works. Next, we use some traditional method to analyze these data. Then, we use some deep learning method to figure out our problem. And, we try to implement our model with more sensible models, and do some ensemble. At last, we talk about our comparison and conclusions.**

***Keywords*—Drug analysis, Machine learning, Deep learning, ECFP**

I. Introduction

Our Project aims to find some relations from medical informations. Since the average price tag for getting a new medicine has risen up to 2.5 billion dollars. And the estimated time that need cost is 10 to 15 years. So it is meaningful if we can get some information before we start the experiments. And hope those information can drive some decision for us.

Furthermore, the number of potential medicines is pretty huge while the number we can try is strictly limited by the time and finance cost.

A. Data Description

The datasets have the target, smile, fingerprint, occp, value and measurement.

Target is the object influenced by the medicine. In the CH dataset we have only one target. As for the D2 dataset, we have many target, which we need to predict with different bucket.

Smile is the formula for the different medicine that is the origin information we have.

Fingerprint is the import information we generate from the structure with technology method. We have 1024 different binary value. From the knowledge of the technology, we can not get the fixed strides from this fingerprint, so the information is pretty hard to extract from it.

OCCP is another import information, that is also 1024 columns. We think this may stands for occupancy, i.e., how many times that each of the 1024 features(certain substructure) occurred in the drug compound. At this time we just take them as something correlate to the fingerprint.

Value is the target we have, and the part we want to use the machine learning to do some regression or classification.

Measurement is the way how we get the value, there are several measurement we used here, at first we take them the same and then put them into different bucket to get a better performance.

II. Related Work

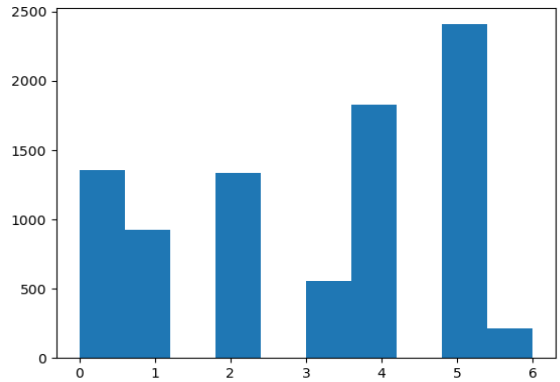
III. Traditional Method

Since the dataset do not obey some order. We split the data simple.

$X\_train, X\_test = X[: int((0.8 * len(X)))], X[int((0.8 * len(X))) :]$

$y\_train, y\_test = y[: int((0.8 * len(X)))], y[int((0.8 * len(X))) :]$

The target distribution seem like this:



A. Evaluation results

We use F1 Score to check performance. In statistical analysis of classification, the F1 score is a pretty good measure of the test evaluation. This macro F1 score including both precision and recall. Precision is the number of true positive divided by all positive results, and recall is the number of true positive divided by the positive results that should have

been returned. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst at 0.

$$F1 = 2 * (precision * recall)/(precision + recall)$$

B. traditional results

We used Support Vector Machine, Random Forest(number of decision tree, depth of decision tree), Naive Bayes to predict the results. And the results are shown below.

Method	SVM	NB	RF(100,100)	RF(100,15)
Train_F1	0.5066	0.4913	0.9935	0.8467
Test_F1	0.1334	0.1362	0.1282	0.1211

TABLE I. TRADITIONAL METHOD F1 SCORE

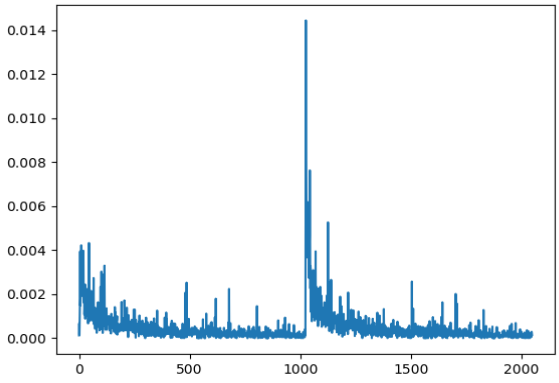
No one has the a better result for our result. Since one divided by seven equals to 0.14285.

C. Random Forest

However, we still looked into the important feature of data. And plot one graph for these 2048 features. And choose the first 25 in fingerprint and the first 25 in occp.

RF accuracy : 0.130714677679 RF train accuracy: 0.982164744178

The result does not change anyway.



RF accuracy : 0.130714677679 RF train accuracy: 0.982164744178

The result does not change anyway.

IV. DEEP LEARNING

A. Basic Neural Network

We will try to use several structures of Basic Neural Network to check the prediction.

1) 1 hidden layer 2048-1024-7:

convergence in 6000 iterations
Training set accuracy:0.6908
Testing set accuracy:0.2520
Cost:1.3206

2) 2 hidden layer 2048-1024-512-7:

convergence in 3200 iterations
Training set accuracy:0.43
Testing set accuracy:0.2485
Cost:1.3959

3) 3 hidden layer 2048-1024-512-256-7:

convergence in 2800 iterations
Training set accuracy:0.9628
Testing set accuracy:0.3042
Cost:1.1936

4) 10 hidden layers:

Training set accuracy:0.99
Testing set accuracy:0.33
Cost:0.99

B. Convolutional Neural Network

We will try to use several structures of CNN to check the prediction.

1) 2-con-2-func: stride : 2 padding: 8

Training set accuracy:0.99
Testing set accuracy:0.33
Cost:0.99

2) 2-con-2-func: stride : 2 padding: 16

Training set accuracy:0.99
Testing set accuracy:0.33
Cost:0.99

3) 2-con-2-func: stride : 4 padding: 8

Training set accuracy:0.99
Testing set accuracy:0.33
Cost:0.99

4) 2-con-2-func: stride : 4 padding: 16

Training set accuracy:0.99
Testing set accuracy:0.33
Cost:0.99

C. Recurrent Neural Network

We will try to use several structures of CNN to check the prediction.

1) structure:

Training set accuracy:0.99
Testing set accuracy:0.33
Cost:0.99

2) structure:

Training set accuracy:0.99
Testing set accuracy:0.33
Cost:0.99

3) structure:

Training set accuracy:0.99
Testing set accuracy:0.33
Cost:0.99

4) structure:

Training set accuracy:0.99
Testing set accuracy:0.33
Cost:0.99

D. Combine Fingerprint and OCCP

RNN Accuracy: 0.26

CNN: Accuracy: 0.31

Basic NN: Accuracy: 0.24

E. choose one using cross validation

F. confusion matrix for the first 100 rows in test set

V. Conclusion

We do not run the neural network more than 10,000 iterations.

Up to now, the best result is from ... . We can implement our model by using ensemble method. Adding more features into the dataset, or adding more instance into the dataset.

Acknowledgment

Thanks to ...

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.