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: husiyuan@gwmail.gwu.edu

Ruoyu Wang
Columbian College of Arts and Sciences
The George Washington University
Email: wry@gwu.edu

***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. At the deep learning part, we improve our accuracy from –%, to –%. 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.

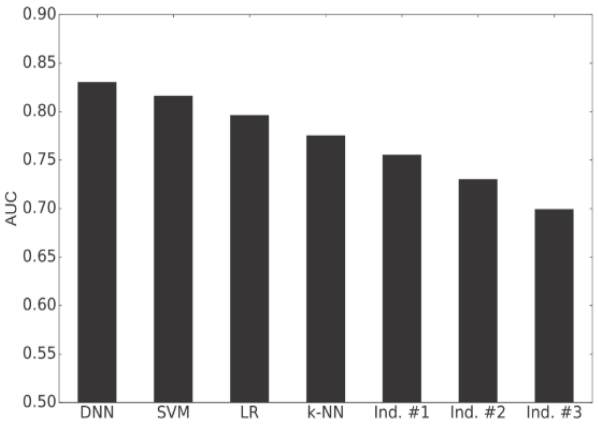
Measurement is the way how we get the value, there are several measurement we used here, at first we take them the same and then put them into different bucket to get a better performance.

II. Related Work

Up to now, deep learning has become the dominant algorithm among many data driven fields. Such as speech recognition, writing recognition, and medical informatics. In 2012, deep learning was introduced by Hinton and his team. And his team win the kaggle competition with their DNN model as a error rate of 16.4 %. This is the first time of deep learning into QSAR in Merck challenge in 2012. While the second team which based on traditional models only achieved the error rate of 26.2%.

In 2014, Dahl used another multi-task neural networks for this QSAR applications based on the kaggle algorithms.

Same year, Hochreiter and co-workers published a peer-review paper. In that essay, the authors used the ChEMBL database including 743,336 compounds, and 5069 targets. They used ECFP4 fingerprints and have a AUC of 0.83.



A. Other related models

Prediction / Competition	DNN Models	Comments	Non-DNN Models	Comments
Merck Kaggle Challenge (Activity)	0.494 R^2	DNN-based model was the top performing model in the competition.[62]	0.488 R^2	Best non-DNN model in the competition.[140]
	0.465 R^2	Median DNN-based model recreated by Merck post-competition.[66]	0.423 R^2	Best non-DNN model (RF-based) by Merck post-competition.[66]
Activity	0.830 AUC	MT-DNN based model trained on the ChEMBL database.[68]	0.816 AUC	Best non-DNN model (SVM) trained on the ChEMBL database.[68]
	0.873 AUC	MT-DNN based model trained on the PCBA database.[70]	0.800 AUC	Best non-DNN model (RF) based model trained on the PCBA database.[70]
	0.841 AUC	MT-DNN based model trained on the MUV database.[70]	0.774 AUC	Best non-DNN model (RF) based model trained on the MUV database.[70]
NIH Tox21 Challenge (Toxicity)	0.846 AUC	DeepTox (MT-DNN based model) was the top performing model.[86]	0.824 AUC	Best non-DNN model (multi-tree ensemble model) was placed 3rd in the Tox21 challenge.[141]
	0.838 AUC	Runner up in Tox21 challenge was based off associative neural networks (ASNN).[142]		
	0.818 AUC	Post-competition MT-DNN model.[70]	0.790 AUC	Post-competition RF model.[70]
Atom-level Reactivity/ Toxicity	0.949 AUC	DNN-based model that predicts site of epoxidation, a proxy for toxicity.[80]	-	No comparable model in the literature that can identify site of reactivity or toxicity.
	0.898 AUC	DNN-based model that predicts site of reactivity to DNA.[84]		
	0.944 AUC	DNN-based model that predicts site of reactivity to protein.[84]		
Protein Contact	36.0% acc.	CMAPpro (DNN-based model).[106]		
	34.1% acc.	DNCON (DNN-based model).[108]	29.7% acc. 28.5% acc.	Best non-DNN model reported in CASP9, ProC_S3 (RF-based model)[28] and SVMcon (SVM-based model)[27] are listed respectively.

In 2015, Merch published a study paper, talking about the comparison between DNNs and RF-based models.

III. Regression

1) introduction:

For these 2 drug datasets, a potential interesting issue is whether there exists relationship between measurement value and fingerprint and or occp. So, several regression methods are tried.

2) Data Preprocessing:

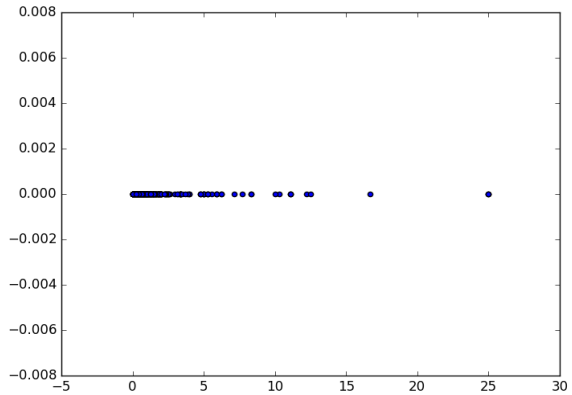
To avoid overfitting due to some very large measurement values, the measurement values are inverted from the original dataset.

3) Random Forest Model:

The model with 4 estimators, R square is 0.42433962429
The model with 1 max-depth, R square is 0.585149652273
The model with 1 max-depth and 4 estimators, R square is -0.036793710793

4) Neural Network with single Neuron:

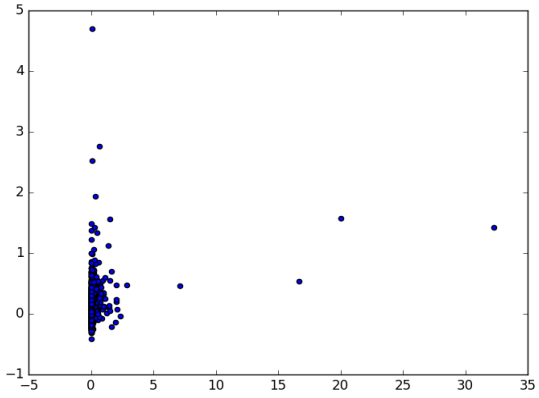
Cost function using MSE



Iteration Times: 10000 MSE: 3714.18 (Actually the MSE of single neuron keeps this value from the first iteration)

5) Convolution Neural Network (Cost Function using MSE):

2 Convolution Layers 1 Max Pool Layer 2 Fully Connected Layers Iteration Times: 1800 MSE: 1.1552



6) conclusion of regression:

All the regression methods tried for this issue show very poor results, so probably it is not a feasible task to build regression model showing relationship between measurement values and fingerprints combining occp, or perhaps the regression issue request another way to organize the data set.

IV. Traditional Method for Classification

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

$$X, y = shuffle(X, y, random\_state = 0)$$

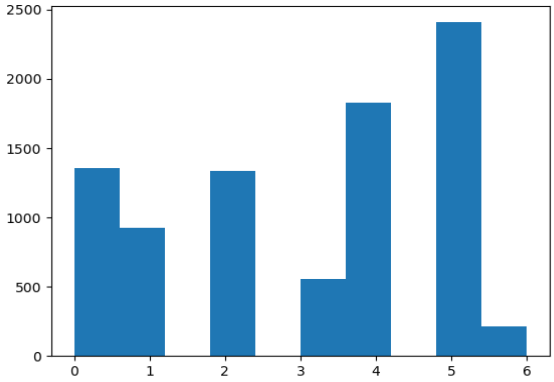
X part:

$$X\_train, X\_valid, X\_test = X[: int((0.6 * len(X)))], X[int((0.6 * len(X))) : int((0.8 * len(X)))], X[int((0.8 * len(X))) :]$$

y:part:

$$y\_train, y\_valid, y\_test = y[: int((0.6 * len(X)))], y[int((0.6 * len(X))) : int((0.8 * len(X)))], y[int((0.8 * len(X))) :]$$

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.

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

The reason why we do not use R square or AUC is that sometimes the increasing of AUC doesn't really reflect a better classifier.

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.

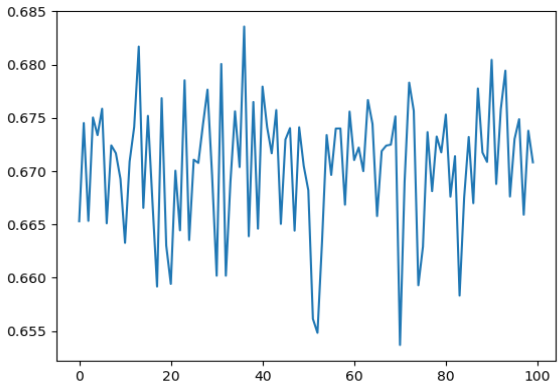
And the result some kind of depends of the shuffle results.

Method	SVM	NB	RF(100,100)	RF(100,50)
Train_F1	0.5005	0.4840	0.9938	0.9938
valid_F1	0.5049	0.4196	0.8023	0.7974
test_F1	0.4740	0.4386	0.7919	0.7859

TABLE I. TRADITIONAL METHOD F1 SCORE

The random forest has a best results among the traditional methods. Since it uses the ensemble methods, and achieve the train set accuracy to the 0.9938.

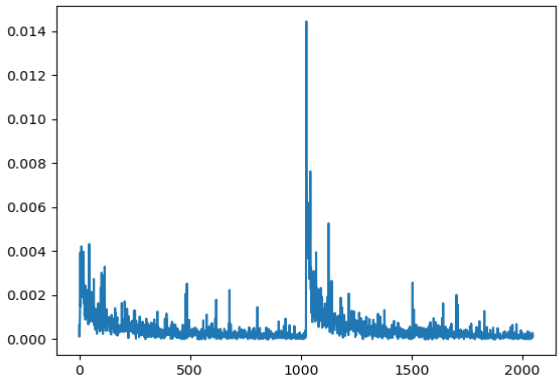
The 100 decision trees for the accuracy:



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.

The result does not change anyway.



Training set accuracy:0.98448
Validation set accuracy: 0.71011
Testing set accuracy:0.7701

The result does not change a lot. So We can assume that the result is gotten most from the first 50 important features. If we want to get some more useful information to improve our results. We need to pay more attention to the last rest part.

V. DEEP LEARNING FOR CLASSIFICATION

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:

Training set accuracy:0.9034
Validation set accuracy: 0.7913
Testing set accuracy:0.7596
Cost:1.22

The train accuracy is not close to the RF results, but the result is better than a single decision tree. So we can assume that it has some space for improvement.

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

Training set accuracy:0.9464
Validation set accuracy: 0.7020
Testing set accuracy:0.6634
Cost:1.22

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

Training set accuracy:0.9775
Validation set accuracy: 0.669565
Testing set accuracy:0.6378
Cost:1.22

For these three experiment, the results are not bad, that maybe we can get some good performance after ensemble. Because the single neural network is better than the single decision tree. With the considering of both the time cost and performance, maybe the ensemble of the just one single layer is better for us.

B. Convolutional Neural Network

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

1) CNN1: stride : 2 padding: 8

iterations 1000
Training set accuracy:0.994783
Testing set accuracy:0.811703
Cost:0.014

Compare with the experiment of Basic neural network, CNN has a faster calculation performance and the nice iteration of convergence.

After 200 iterations , the test accuracy has reached up to 0.73. And after 1000 iterations, the test accuracy has reached up to 0.81. Furthermore, comparing with the random forest results. The single convolution neural network has the same level experience. So we hold this one as the ensemble method into the wait list.

2) CNN2: stride : 2 padding: 16

After 600 iterations the accuracy has been reached 0.80.

iterations 1000
Training set accuracy:0.994059
Testing set accuracy:0.80591
Cost:0.02

3) CNN3: stride : 4 padding: 16

iterations 1000
Training set accuracy:0.9942
Testing set accuracy:0.7931
Cost:0.024

4) CNN4: stride : 8 padding: 32

iterations 1700
Training set accuracy:0.994783
Testing set accuracy:0.775782
Cost:0.012

A large padding is not useful for this dataset. And better performance prefer less padding and stride.

C. Recurrent Neural Network

As the time we take the finger print as the part of language in the medical fields. RNN may have some helpful in this classification.

1) 1024-512-256:

Just 100 iterations , the train accuracy has reached up to 0.9945, while the test accuracy only 0.4710. Furthermore, the test accuracy increase very slowly with the steady train accuracy. Although the result well be better, we do not choose it considering the time cost.

iterations 7*100
Training set accuracy:0.99478334
Testing set accuracy:0.47914
Cost:0.80504

2) 512-256-128:

iterations 79*100
Training set accuracy:0.9944535
Testing set accuracy:0.0.6292
Cost:0.1.1750

We did not run the total part of RNN until it is convergent. Because it cost so much for us. If we have the calculation power for this part the result is better than the single decision tree.

D. Combine Fingerprint and OCCP

We assume that there are some relation between fingerprint and occp. So we choose several models to experiment.

1) CNN: CNN with the first layer padding equals 2 and stride equals 2. After 600 iterations the test accuracy reaches 0.80. Compared with the first CNN model(test accuracy 0.811703), this one has a test accuracy as 0.80591 after 1000 iterations.

iterations 2000
Training set accuracy:0.994349
Testing set accuracy:0.808227
Cost:0.010687

2) RNN:

iterations 100*100
Training set accuracy:0.99420375
Testing set accuracy:0.2346
Cost:0.9869

VI. Conclusion

A. summary

In our project, we experiment several neural works with different parameters. Aims to analyze those information to get a direction of model adjustment. Such as :
1)the number of iterations which the loss function has been convergent.
2) The accuracy of training set
3) The accuracy of testing set
4) The loss function after convergent

B. implement and future work

We do not run the neural network more than 10,000 iterations. Since the cost is extremely huge. And the dataset is not big enough for our analysis.

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. And using cross-validation as well.

The most import thing we need to implement is feature engineer. In this project we still not find a good way to extract information from the fingerprints and occp.

Acknowledgment

Thanks to our professor Zeng Chen who give us many help on our project.

References

[1] Deep Learning for Computational Chemistry,Garrett B. Goh,Nathan O., Hodas, Abhinav Vishnu

[2] Ajay; Walters, W.p.;Murcko, M.A.J.Med.Chem. 1998,31,3314

[3] Burden, F.R.; Ford, M.G.; Whitley, D.C.; Winkler, D.a.J.Chen=m. inf. Comput. Sci. 2000, 40,1423

[4] Cramatica, P.QSAR Comb. Sci. 2007, 26, 694

[5] Verma, J.; Khedkar, V.M.; Coutinho, E.C.Curr. Top. Med Chem. 2010, 10,95

[6] Kaggle

[7] Dahl, G.E.; Jaitly, N.; Salakhutdinov, R.arXin:1406.1231 2014

[8] Ma,J.;Sheridan, R.P.; Liaw, A.;Dahl, G.E.; Svetnik, V.J. Chem. Inf. Model. 2015m 55m 263