VIRTUALMAN www.virtualman.ai contact@virtualman.ai

Training strategy for unbalanced small datasets in deep learning

Hsin Liu [a,c], Cheng-Wei Lee [a], Bo-Han Su [b], and Yufeng J. Tseng [b,c,d,e]

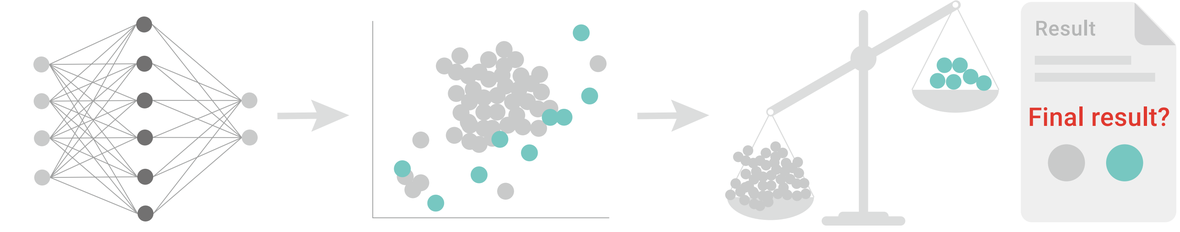
a. Virtual Man Inc. 3F., No. 196, Sec. 2, Fuxing S. Road, Da'an Dist., Taipei 106, Taiwan
b. Department of Computer Science and Information Engineering, National Taiwan University, No. 1 Sec. 4, Roosevelt Road, Taipei 106, Taiwan.
c. Graduate Institute of Biomedical Engineering and Bioinformatics, National Taiwan University, No. 1 Sec. 4, Roosevelt Road, Taipei 106, Taiwan.
d. The Metabolomics Core Laboratory, Center of Genomic Medicine, National Taiwan University, No. 1 Sec. 1, Jen Ai Road, 106, Taiwan
e. Drug Research Center, School of Pharmacy, College of Medicine, National Taiwan University, No. 1 Sec. 4, Roosevelt Road, Taipei 106, Taiwan.

Abstract

Big datasets have been keys to deep learning and the neural network approach applied to them in the past few years. However, one never has the luxury in medicinal chemistry compared to image processing field where large accessible data were readily available. The smaller datasets are often due to the lack of published experimental results which might be affected by including complex experimental design, expensive experimentation, or simply limitations in techniques. Also, the nature of medicinal chemistry chasing after more active compounds make almost published data unbalanced—that is having few positive data with mostly negative data. It would be invaluable to be able to train a model with unbalanced small dataset in medicinal chemistry for drug development in particular. In this work, we proposed a training strategy for unbalanced small datasets. The strategy includes selecting the sampling ratio, core deep learning methods, fingerprint selection, and descriptor merge of fingerprint and automatic feature extraction by deep learning. We chose the Ames test for mutagenicity as the example in this study due to its available information for validation study; and also the entire dataset could be divided in segments to simulate unbalanced small datasets for training and discussion. Overall, the up-sampling method is able to rebalance the data distribution in different categories and demonstrates better performance in both convergence speed and balanced accuracy.

Introduction

For model training, the data adjustment strategy and design of neural networks were the two major focuses other than adjusting the parameters relating to the model itself. It is especially important when it comes to the end-to-end model training of neural network while the features selection is no longer necessary. As such, the focus is on how to manage the unbalanced small dataset. In this work, we demonstrate our data adjustment strategy on the mutagenicity dataset which is a well-recognized dataset for various studies. Moreover, we had adopted graph convolutional neural network (GCN) as the class of neural networks. The simulated unbalanced dataset was created by randomly sampling the original dataset, and part of the samples was used as the test data.



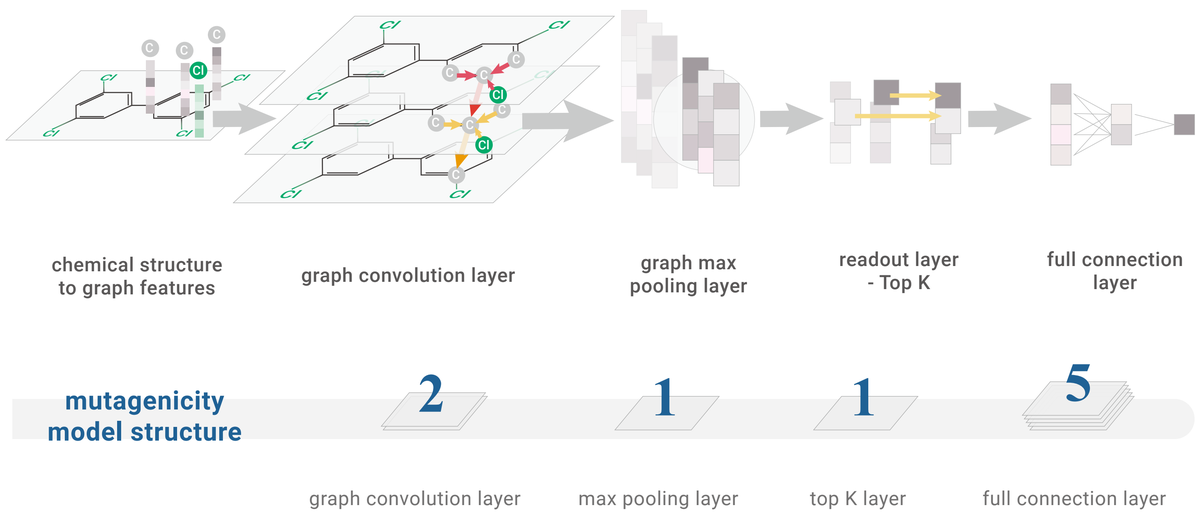
Material & Methods

Dataset Retrieval & Curation

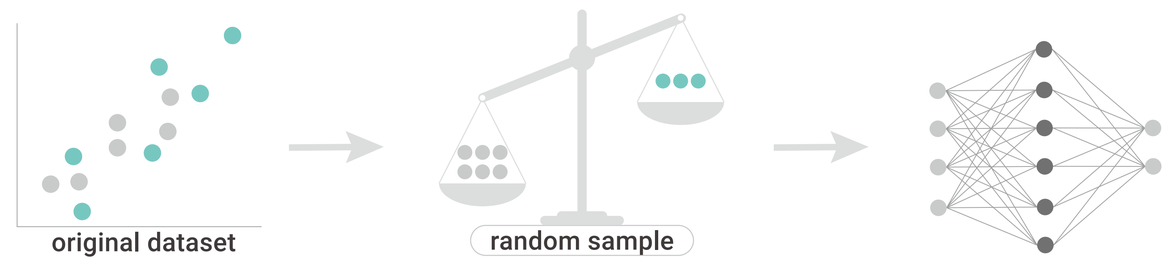
Mutagenicity database (Kuo-Hsiang Hsu 2016) was used. In this data set, both positive and negative data have reached the scale of more than 3,000. This allowed sufficient data to be partitioned to simulate the unbalanced data in this study.

Methods

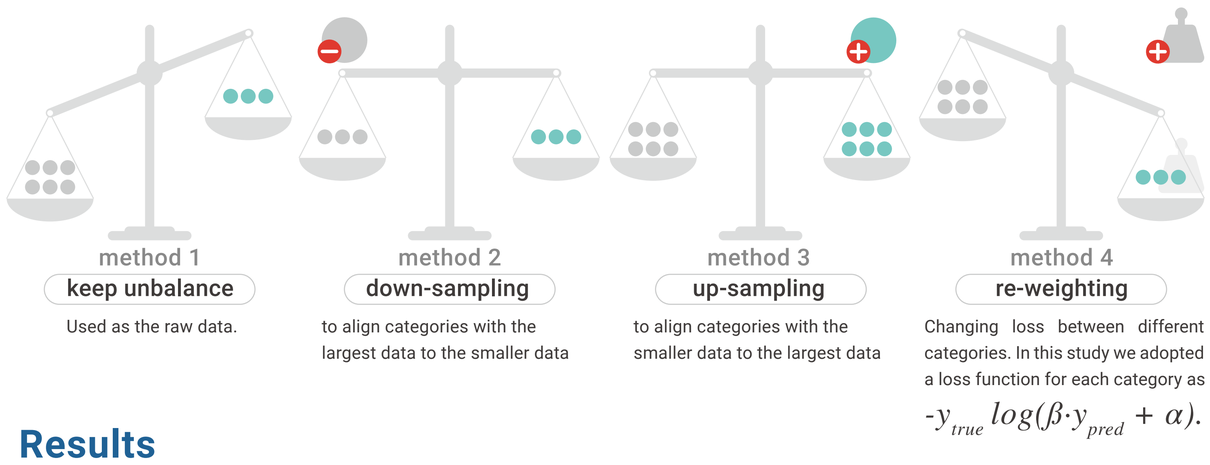
Graph Convolutional Neural Network (GCN)[1] was used in this study as the prediction method. The molecular structure was transformed into graph structure as model input, and the atom was encoded as 77 features.



Simulating unbalance data was performed by randomly sampling the complete data, and leave some as the test data



With the above dataset, this study applied the following four methods to adjust the unbalanced data and compared with each dataset for the performance.



Results

The followings indicated the results with statistically significance:

dataset #2 ~ dataset #5 randomly sample from dataset #1 to 250 positive samples and 2500 negative samples.

dataset #1: original dataset

train	truth: true	truth: false
prediction: true	1978	266
prediction: false	251	1985
	0.89 Sen.	0.88 Spec. / 0.88 Acc.

valid	truth: true	truth: false
prediction: true	202	56
prediction: false	57	185
	0.78 Sen.	0.77 Spec. / 0.77 Acc.

test	truth: true	truth: false
prediction: true	346	99
prediction: false	98	357
	0.78 Sen.	0.78 Spec. / 0.77 Acc.

dataset #2: adjusted by method 1

train	truth: true	truth: false
prediction: true	109	20
prediction: false	88	2023
	0.55 Sen.	0.99 Spec. / 0.95 Acc.

valid	truth: true	truth: false
prediction: true	7	3
prediction: false	25	240
	0.22 Sen.	0.99 Spec. / 0.90 Acc.

test	truth: true	truth: false
prediction: true	120	9
prediction: false	324	447
	0.27 Sen.	0.98 Spec. / 0.63 Acc.

dataset #3: adjusted by method 2

train	truth: true	truth: false
prediction: true	214	10
prediction: false	4	220
	0.98 Sen.	0.96 Spec. / 0.97 Acc.

valid	truth: true	truth: false
prediction: true	21	3
prediction: false	10	16
	0.68 Sen.	0.84 Spec. / 0.74 Acc.

test	truth: true	truth: false
prediction: true	256	88
prediction: false	188	368
	0.58 Sen.	0.81 Spec. / 0.69 Acc.

dataset #4: adjusted by method 3

train	truth: true	truth: false
prediction: true	2116	63
prediction: false	2	2139
	1.00 Sen.	0.97 Spec. / 0.98 Acc.

valid	truth: true	truth: false
prediction: true	10	12
prediction: false	22	231
	0.31 Sen.	0.95 Spec. / 0.88 Acc.

test	truth: true	truth: false
prediction: true	192	29
prediction: false	252	427
	0.43 Sen.	0.94 Spec. / 0.68 Acc.

dataset #5: adjusted by method 3 and method 4 with $\alpha=0.1$ and $\beta=0.9$

train	truth: true	truth: false
prediction: true	1938	356
prediction: false	106	1760
	0.95 Sen.	0.83 Spec. / 0.89 Acc.

valid	truth: true	truth: false
prediction: true	22	41
prediction: false	10	202
	0.69 Sen.	0.83 Spec. / 0.81 Acc.

test	truth: true	truth: false
prediction: true	280	99
prediction: false	164	357
	0.63 Sen.	0.78 Spec. / 0.71 Acc.

Conclusion

- The down-sampling method in theory could rebalance the data distribution. However, as the total data information dropped, the final results tended to be compromised.
- The up-sampling method in theory could rebalance the data distribution. However, the final model performance was not improved and results seemed to indicate unbalance.
- Among the four data adjustment methods, the combination of up-sampling and re-weighting gave the best model performance.

References

1. David K. Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alan´ Aspuru-Guzik, and Ryan P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In Advances in neural information processing systems (NIPS), pp. 2224–2232, 2015.