Improved Label Propagation Model to Predict Drug - drug Interactions

Zhice Yan, Lasheng Zhao, Xiaopeng Wei and Qiang Zhang^a

Advanced Design and Intelligent Computing Laboratory, Dalian University, Dalian, China

Abstract. Drug-drug interactions (DDIs) is one of the most concerned issues in drug design. Accurate prediction of potential DDIs in clinical trials can reduce the occurrence of side effects in real life of drugs. Therefore, we propose a model to predict DDIs. The model integrates several methods that can improve label propagation algorithm. Firstly, the chi-square test (CHI) method is adopted to filter or select the features that contain a large amount of information. Secondly, the sample similarity calculation method is reconstructed by label similarity and feature similarity. Then the label initialization information of unlabeled samples is constructed. Finally, we use label propagation algorithm to estimate the labels of the unlabeled drugs. The results show that the proposed model can obtain higher the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPR), which provides a favorable guarantee for the discovery of DDIs in the clinical stage.

1 INTRODUCTION

Drugs may interact when more than one drugs are co-prescribed. Drug-drug interactions (DDIs) may cause serious side effects, which can lead to other diseases or patient's death [1-3]. In order to prevent the unsafe use of new drugs, they must be tested in clinical trials or be allowed to appear on the market after phase IV clinical trials. Although a large number of DDIs are found during the clinical trials, there are still many DDIs on the market. According to related reports abroad, the probability of adverse drug reactions in hospitalized patients is 10% to 20%, and of which 0.24% to 2.9% of patients died from adverse drug reactions. Therefore, the effective detection of potential DDIs of new drugs is an urgent problem to be solved in clinical trials.

In recent years, Data Mining and Machine Learning are the mainstream of medical big data research. In the field of drug research, a great number of Data Mining methods and Machine Learning methods were also proposed to predict DDIs of new drugs. The existing methods are roughly divided into two types: similarity-based methods and classification-based methods. The similarity-based methods used the hypothesis that similar drugs may generate same effects with a same drug. Vilar et al. [4] proposed the similarity-based prediction method and modeled interaction profile fingerprints to predict DDIs [1]. Fokoue et al. [5] developed the similarity calculation method between two drug pairs, and built a Tiresias framework of multiple drug data sources to predict the DDIs of different sources. Zhang et al. [6] used three representative methods: the matrix perturbation method, the neighbor recommender method and the random walk method to predict potential DDIs

by integrating chemical, biological, phenotypic, and network data. Celebi et al. [7] adopted Rooted PageRank algorithm to establish the recommended model by drug treatment, similarity of genes, phenotypes and chemical substructures. Zhang et al. [8] applied label propagation method to predict DDIs based on drug side effects, drug off side effects and drug chemical substructures. Classification-based methods predict the DDIs as binary classification tasks. Cheng et al. [9] proposed a heterogeneous Network Aided Reasoning (HNAI) framework to assist DDIs prediction, and the framework employed five predictive models: naive Bias, decision tree, k-nearest neighbor, logistic regression and support vector machine. Jamal et al. [10] used the Weka tool to filter neurodrug data, and established a predictive model of neurodrugs based on the logistic regression algorithm. Yang et al. [11] developed multitask dyadic regression to predict each specific DDIs type of all drugs.

Despite their initial success, there are still many DDIs not detected or observed in clinical trials. On the one hand, it is difficult to acquire DDIs data, and the DDIs data is usually mastered and studied by large companies or organizations. On the other hand, the similarity-based methods directly use Tanimoto Coefficient (TC) to compute similarities between all the fingerprints [6, 8], but ignore the importance of drug features and the connection of multi-label drugs. The method of calculating similarities can't make the drugs clearly separated (there are a lot of overlaps in drug prediction results). The method reduces the accuracy of the drug prediction.

In this paper, we considered the lack of similarity calculation methods in drug research, and Zhang et al. [8] proposed label propagation algorithm to predict DDIs

^aCorresponding author: azhangq26@126.com

based on the similarity-based methods. Thus we proposed several methods to improve the label propagation model. These methods include: the CHI feature filtering method, the Laplace operator method, the Label similarity method and the Unlabeled sample label initialization method. According to performances of prediction models, we evaluated the methods.

The rest of the paper is organized as follows. Section 2 introduces the label propagation algorithm [8]. While in Section 3 we will be ready to introduce several improved methods based on the label propagation model. Next we evaluate their performances with real drug data in Section 4. Finally we conclude our work and look forward to future work in Section 5.

2 Label Propagation Model

To present our method, in the section, we will introduce the two main blocks: Label propagation algorithm and similarity measure.

2.1 Label propagation algorithm

Label propagation algorithm is a semi-supervised learning algorithm which mainly addresses the following problem: given an undirected weighted network with n nodes where a small portion of them are labeled, spread the labels to the rest unlabeled nodes [8, 12]. In this network, they used different drugs as nodes, and computed the edge weights with drug similarities. For each drug, all other drugs were labeled as positive if they are known to have DDIs with this drug, and using label propagation algorithm to spread the labels to unlabeled nodes on the network example in Figure 1.

Figure 1. Drug-drug interactions network. The right picture can be observed that Drug 1 is connected with Drug 7, Drug 11 and Drug 18 as well as Drug 1 isn't connected with Drug 2. The connection shows that there are DDIs between them.

To better depict this idea, the algorithm uses TC to compute the similarities between the all drugs, and obtains an $n \times n$ affinity symmetric matrix A . For all drugs, the algorithm constructs a label matrix Y , where $Y_{ij}=1$ if drug i is known to have DDI with drug j , and $Y_{ij}=0$ otherwise. To ensure convergence of the updates, the original affinity matrix A needs to be normalized. In the study of Zhang et al. [8], they used Bregmanian Bi-Stochasticity (BBS) algorithm [13] to normalized the similarity matrix A and denoted the normalized matrix as W . Finally, the label propagation algorithm was used

to spread the labels to unlabeled nodes through iterations again and again in the network. Scores of unlabeled nodes can be obtained by

$$F = (I - u)(I - uW)^{-1}Y \quad (1)$$

Here, u represent each drug node "absorbs" a portion u of the label information from its neighborhood, and retains a portion $1 - u$ of its initial label information.

2.2 Similarity measures

The prediction of drug side effects usually uses TC, also known as the Jaccard index, to compute similarities between all the chemical structure drugs. The TC between drug A and drug B is defined the ratio between the numbers of features in the intersection to the union of both:

$$TC(A, B) = \frac{A \cap B}{A \cup B} \quad (2)$$

Then we get a symmetric matrix A so that the rows and columns represent drugs and A_{ij} represent the similarity between drug i and drug j .

3 Proposed Methods

We found that the prediction F have a strong relationship with the normalized matrix W and the initialized label matrix Y in formula (1), where matrix W is normalized by the drug similarity matrix A . Therefore, the improvement of the label propagation algorithm is the improvement of Similarity measures and Initialization Label type. We proposed several methods: Feature Select method, Improved Similarity methods (Laplace Operator method and the Label Similarity method) and Unlabeled Sample Label Initialization method. As shown in Figure 2.

Figure 2. Improved Label Propagation Framework

3.1 Feature Select Method

It is different that each dimension of drug molecule carries information. In our study, the initial data of drug chemical structure was preprocessed by deleting the chemical feature columns whose feature values are all 0. Then we used the chi-square test (CHI) method to filter the preprocessed features, and chose the features that are greater than the threshold we set.

The CHI method measures the strength of the relationship between feature t_j and category C_k by the assumptions that feature t_j and category C_k satisfy the χ^2 distribution of a dielectric degree of freedom. If the

correlation between the feature t_j and the category C_i is higher, χ^2 is also the higher. Then the feature is selected based on the basis. In the study, we used the improved CHI method proposed by Gao et al. [14] to select the chemical features of drugs.

Table 1. Relationship between features and categories

Category features	Category	Other categories	Sum
The number of t_i	C_k a	b	a+b
The number of without t_i	c	d	c+d
Total Number	a+c	b+d	N=a+b+c+d

The deviations of feature t_j and categories C_k are:

$$\chi^2(t_i, C_k) = \lambda_k \times \eta_k \times \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \quad (3)$$

where λ_k represents the frequency of the feature t_j appears in the class C_k , and its formula is

$$\lambda_k = \frac{a}{a+c} \quad (4)$$

In the calculation results, the greater the value of λ_k shows the better performance of the feature t_j , in other words, the higher classification ability.

η_k represents the degree that the feature appears in a certain category, which can be expressed as

$$\eta_{ki} = \frac{\frac{a}{a+c+1}}{\frac{b}{b+d}+1} \quad (5)$$

In formula (5), adding 1 to prevent the numerator and the denominator are 0.

3.2 Improved Similarity Methods

Since there is no complete dissimilarity between two different drugs in real life, a certain connection exists between them more or less. In the similarity matrix of chemical structure, there is a case that the similarities are 0 between a drug and other drugs. However, It is unreasonable to set the similarities are 0 between the drug and other drugs, because the drug have a same effect with the other drugs whose similarities are 0 in the labels matrix. In this study, we drew on the method of Laplace smooth change and proposed the similarity method is adjusted by the Laplace Operator. The improvement formula is as follows

$$TC(A, B) = \frac{|A \cap B| + 1}{|A \cup B| + 1} \quad (6)$$

We considered the case that Chemical structure of the drugs can not completely express all the characteristics of the drugs and the labels also impact on the characteristics of the drugs. It may be possible that the combination of drug label similarity and chemical

structure similarity can improve the prediction accuracy of DDIs. In this section, we proposed a new similarity method to compute the similarities between pair of drugs. The similarity is calculated [15] as follows

$$S(i, j) = A(i, j) \times C(i, j) \quad (7)$$

where S represents new drug similarity, and C represents the similarity between label and label of drugs. C can be obtained as follows

$$C(i, j) = \begin{cases} 1 - \frac{1}{\sum_{t=1}^m q(t)} \sum_{t=1}^m q(t) |l'_i - l'_j|; & \forall i \leq N, j \leq N \\ \frac{1}{|N_{Lj}|} \sum_{x_k \in N_{Lj}} \left| 1 - \frac{1}{\sum_{t=1}^m q(t)} \sum_{t=1}^m q(t) |l'_i - l'_k| \right|; & \forall i \leq N, N \leq j \leq N + N', N_{Lj} \neq \emptyset \\ \frac{1}{|N_{Li}|} \sum_{x_k \in N_{Li}} \left| 1 - \frac{1}{\sum_{t=1}^m q(t)} \sum_{t=1}^m q(t) |l'_i - l'_k| \right|; & \forall j \leq N, N \leq i \leq N + N', N_{Li} \neq \emptyset \\ \overline{C(i, j)}; & \text{others} \end{cases} \quad (8)$$

In formula (8), N_{Lj} represents a subset of the $k - \xi$ nearest neighbor set of unlabeled drug j , and $\overline{C(i, j)}$ represents the mean from the three previous cases. When N_{Lj} or N_{Li} or the sample is the samples to be tested, we set the similarities to $\overline{C(i, j)}$. In this formula, we use $q(t)$ represents the weight of the t -th label and which can be rewritten as

$$q(t) = 1 + \ln\left(\frac{N_p}{N_t}\right) \quad (9)$$

where N_p represents the number of total samples, and N_t represents the number with the t -th label in the training set. The weight method reduces the importance of the strong label and enhances the importance of the weak label.

3.3 Feature Select Method

In formula (1), we make $H = (I - u)(I - uW)^{-1}$, then formula (1) can be written as

$$\begin{bmatrix} F_l \\ F_u \end{bmatrix} = \begin{bmatrix} H_{ll} & H_{lu} \\ H_{ul} & H_{uu} \end{bmatrix} \times \begin{bmatrix} Y_l \\ Y_u \end{bmatrix} \quad (10)$$

For the convenience of observing the relationship between F and Y_u , we first rewrote the objective of formula (10) as

$$\begin{cases} F_l = H_{ll} Y_l + H_{lu} Y_u \\ F_u = H_{ul} Y_l + H_{uu} Y_u \end{cases} \quad (11)$$

It can be clearly observed that Y_u affect F_l and F_u . Therefore, the study of Y_u may increase the accuracy of the final prediction. In this paper, we used the threshold 0.5 to divide the drugs into similarities or dissimilarities, and calculated the probabilities P_{diff} of the dissimilar

drugs interaction and the probabilities P_{sim} of similar drugs interaction in the whole training drug dataset, respectively. Finally, we given the unlabeled drugs the probabilities of the initial labels as follows

$$\mathbf{Y}_u(i, j) = \begin{cases} P_{diff}, \mathbf{sim}(i, j) < 0.5 \\ P_{sim}, \mathbf{sim}(i, j) \geq 0.5 \end{cases} \quad (12)$$

4 Experiments

In this section, we evaluated the performances of our methods with experiment on real drug data and discussed the results.

4.1 Datasets

FAERS DDI database. FDA Adverse Event Reporting System (FAERS) contains information on adverse events submitted to FDA, which is designed to support FDA's post-marketing safety surveillance program for drugs and therapeutic biological products. Mined from FAERS, TWOSIDES dataset [16] only contains side effects for pairs of drugs. In this study, we used the unsafe co-prescriptions from TWOSIDES as known set of DDIs. There are 645 drugs and 63,473 distinct pairwise DDIs in the dataset.

Chemical structure dataset. We used PubChem [17] substructure fingerprint to obtain drug chemical features. Each drug was represented by an 881-dimensional binary profile whose elements encode for the presence or absence of each PubChem substructure by 1 or 0, respectively.

4.2 Evaluations

To ensure the validity of the test cases, we carried out all the DDIs associated with a fixed percentage of the drugs, rather than holding out DDIs directly. To be specific, we randomly selected a fixed percentage (20%) of drugs for testing, and moved all DDIs associated with these drugs as the testing set. Then we used the remaining DDIs to construct the models. The model parameters were tuned with cross validation based on the training set. For each testing, we adopted the area under the receiver operating characteristic curve (AUROC) and the area under the

precision-recall curve (AUPR) to report the performances, and computed the mean and the standard deviation to evaluate the methods.

In the experiment, we compared several DDIs prediction methods: (1) Label Propagation with Chemical Similarity (LP_Chemical) [8]. The model used that labeled samples iteratively spread their label information to unlabeled samples by weigh between two nodes in the network. In the Label Propagation model, best parameter we chose is $u=0.95$. (2) Label Propagation with Feature Extraction (LP+CHI). Here, we used the label propagation algorithm to predict DDIs based on the features filtered by the CHI method. And the parameters we optimized is $u=0.95$, $threshold=0.0001$. (3) Label Propagation with Laplace operator (LP+Laplace). In this process, we adjusted the drug similarity through the Laplace operator to establish the prediction model of the DDIs, where $u=0.95$. (4) Label Propagation with Label similarity (LP+Label_Sim). Considering the influence of multi-label in drug similarity, we combined label similarity and feature similarity here, and made the combined results as new similarity, which to predict DDIs by label propagation, where $u=0.96$, $k=2$, $\xi=0.80$. (5) Label Propagation with Initialize label (LP+ $\mathbf{Y}_u$). We initialized unlabeled samples and established a label propagation model to predict DDIs. In this model, when $u=0.95$, the predictive effect is optimal. (6) Label Propagation by Integrating All Ways Improved (LP+CHI+Laplace+Label_Sim+ $\mathbf{Y}_u$). Predicting novel DDIs by integrating label propagation processes of multiple improved methods. In this study, we integrated networks derived from Feature selection, Laplace operator, Label similarity and Initialize label. And when $k=2$, $\xi=0.80$ and $u=0.97$, the prediction accuracy of the model reaches the best.

4.3 Experiment Results

We compared several DDIs prediction models from Tab2.

Table 2. Comparison of DDI prediction methods according to AUROC and AUPR

Methods	AUROC	AUPR
LP_Chemical [8]	0.8063+/-0.0050	0.6457+/-0.0154
LP+CHI	0.8105+/-0.0052	0.6452+/-0.0163
LP+Laplace	0.8113+/-0.0050	0.6458+/-0.0155
LP+Label_Sim	0.8129+/-0.0059	0.6502+/-0.0147
LP+ $\mathbf{Y}_u$	0.8107+/-0.0051	0.6463+/-0.0163
LP+CHI+Laplace+Label_sim+ $\mathbf{Y}_u$	0.8119+/-0.0054	0.6522+/-0.0163

And clear observations: (1) We proposed several improved models that are better than the Label Propagation model [8]. And from the similarity method, LP+Label_Sim method obtained much higher AUROC scores than Laplace Operator (e.g., LP+Label_Sim method achieved averaged AUROC of 0.8129 and AUPR of 0.6502; LP+Laplace method achieved averaged AUROC of 0.8113 and AUPR of 0.6458). (2) It is very obvious that LP+YU are better than LP_Chemical [8] in both AUROC and AUPR. The unknown label initialization method is also a source that can increase the predictive score of label propagation. (3) Our proposed LP + CHI + Laplace + Label_sim + Y_u method combines the improved methods of label similarity and the method of unknown label initialization. Although in our experiments, the LP+Label_Sim method performs better than our LP+CHI+Laplace+Label_sim+ Y_u method in AUROC, but in AUPR aspect, LP+CHI+Laplace+Label_Sim+ Y_u method can show greater value (e.g., In medicine, mistaken predictive drugs do not have a DDI cost more than the cost of the wrong predictive medicine with DDI).

4.4 Discussion

In this study, our methods have achieved good results in DDIs prediction, but there are still some shortcomings:

- In this study, the drug chemical structure data we used downloaded from PubChem [17] are a binary profile, which may have lost some information of the original data. This lead to the inaccuracies in drug similarity as well as the reduction in the accuracy of the DDI prediction.

- In the experiment, we got only 645 data. Compared with the medical big data, the data we used is somewhat insufficient. And we used the DDIs dataset from TWOSIDES. However, TWOSIDES is directly derived from FAERS, which contains some false positives. When using the DDIs dataset, there may be some false positive data, which has an impact on our experimental results.

- In this study, we proposed several methods to improve the accuracy of the model, but without providing the reasons of drug-drug interactions. In future work, we may explore the reasons for the two drug reactions.

5 Conclusion

In summary, we presented several label propagation models to predict adverse drug-drug interactions. Our methods solve the problems of the lack of similarity calculation method and the lack of label propagation method in the field of drug research. And our methods are more accurate in the real dataset, which provides a

favorable guarantee for clinical detection of DDIs, and also achieves the purpose of "safe medication".

Future we will discuss the reasons of adverse drug-drug interactions, and find out which chemical structures causes the adverse effects of the two drugs. This work provides high-precision safety for drug research and development.

ACKNOWLEDGMENTS

We acknowledge the support from the National Natural Science Foundation of China (No.91546123), Program for Changjiang Scholars and Innovative Research Team in University (No. IRT15R07).

References

1. S. Vilar, E. Uriarte, L. Santana, NP. Tatonetti, C. Friedman. PLOS. ONE **8**(2013)
2. T. Prueksaritanont, X. Chu, C. Gibson, D. Cui, KL. Yee. AAPS. J **15**, 629(2013)
3. H. Kusuhara. Drug. Metab. Pharmacokinet **29**, 227(2014)
4. S. Vilar, R. Harpaz, E. Uriarte, L. Santana, R Rabadan. J AM. MED. INFORM. ASSN **19**, 1066(2012)
5. I. Abdelaziz, A. Fokoue, O. Hassanzadeh, P. Zhang, M. Sadoghi. S. S. C. I **104**(2017)
6. W. Zhang, Y. Chen, F. Liu, F. Luo, G. Tian. BMC. BIOINFORMATICS **18**, 18(2017)
7. R. Çelebi, V. Mostafapour, E. Yasar, Özgür. Gümüş, O. Dikenelli. D. E. X. A **14**(2015)
8. P. Zhang, F. Wang, J. Hu, R. Sorrentino. Scientific. Reports **5**, 12339(2015)
9. F. Cheng, Z. Zhao. J. AM. MED. INFORM. ASSN **21**, 278(2014)
10. S. Jamal, S. Goyal, A. Shanker, A. Grover. Scientific. Reports **7**(2017)
11. B. Jin, H. Yang, C. Xiao, P. Zhang, X. Wei, W. Wang. A. A. A. I. (to be published)
12. H. Peng, D. Zhao, L. Li, J. Lu, J. Han, S. Wu. PHYSICA. A **460**, 98(2016)
13. F. Wang, P. Li, A. C. König, M. Wan. Knowledge. and information. Systems **32**, 351(2012)
14. B. Gao, Z. Zhou, W. Yang, Z. Xiao. C. N. K. I. (to be published)
15. X. Bu, X. Cao, H. Gu, H. Journal of Dalian University. (2017)
16. NP. Tatonetti, PP. Ye, R. Daneshjou, RB. Altman. SCIENCE TRANSLATIONAL MEDICINE **4**(2012)
17. W. Yanli, X. Jewen, T. O. Suzek, Z. Jian, W. Jiyao, H. B. Stephen. NUCLEIC. ACIDS. RES. **67**, 623(2009)