A Label Extended Semi-supervised Learning Method for Drug-target Interaction Prediction

JIE Zhao[1, a] and ZHI Cao[1]

[1]College of computer science and electronic engineering, Hunan university, Changsha, 410082, China

[a]scholar_james@163.com

Keywords: Data mining; Drug-target prediction; Label extension; Semi-supervised learning.

Abstract. Computational methods for predicting the new drug-target interactions are more efficient that those experimental methods. Many machine learning based methods have been proposed but most of them suffer from false negative problem. In this paper we extend the original label matrix and adopt weighted lose function to overcome the traditional false negative problem and then propose a label extended semi-supervised learning method called LESSL for drug-target prediction. In our experiment we use two kinds of cross-validation.The results show that our method can raise AUC average by 0.03 and raise AUPR average by 0.04. At last we use the whole dataset as a training set and predict over 10 new drug-target interactions.To conclude our method is efficient and practicable.

Introduction

It is a key point to identify the potential interaction between drugs and targets in modern drug researches. Finding out new targets of a certain drug, we can make a new kind of drug[1] or make drug repositioning[2,3]. Four kinds of interaction data are available including Nuclear Receptors, Enzymes, G-protein receptors and Ion Channels[4] and we can get these data from several public databases such as DrugBank, SuperTarget, KEGG BRITE.

The experimental ways of finding new interaction called in vitro methods are much time-consuming and costly also with some disadvantages[5,6].With the development of computer technology a lot of in silicon prediction ways are proposed including text mining, docking simulation and statistic based machine learning which all give more useful information to in vitro methods. Text mining approaches are short in information redundancy[7]. Docking simulation approaches are sensitive to the structure of targets, but the 3D structures of may targets are still unknown[8].Text mining and docking simulation are also time-consuming thus they can't be used to large-scale prediction while the statistic based machine learning approaches are more effective in large-scale data.

DBSI and TBSI are two typical complex network based methods. DBSI is similar to the item-based recommendation algorithms[9] and TBSI is similar to the user-based recommendation algorithms[10].Different from DBSI and TBSI, the NBI method uses the topology similarity whose theory is just like mass diffusion in physics[11].Then a lot of methods similar to NBI are proposed which only change the way to make diffusion[12].

Another part of learning methods are similarity based approaches. These methods are all under the assumption that similar drugs are more likely to interact with similar proteins. Three kinds of information are needed: the similarity of all the drugs, the similarity of all the proteins and the known interaction between drugs and proteins. Yamanishi et al[13] introduce a supervised method to predict possible drug-protein interactions using bipartite local models. Then Yamanishi et al[14] make further research in pharmacological space and propose a new prediction method. Jian et al[15] improve the BLM model. They introduce the interaction information of the neighbors and propose the BLM-NII model achieved better prediction performance than Yamanishi. Masataka et al[16] use the side effect of the drugs and reconstruct the similarity of all the drugs then use pairwise kernel regression to make prediction. From the semi-supervised learning view XiaZ et el[17] use Laplacian regularized least squares(LapRLS) and net Laplacian regularized least squares(NetLapRLS) models to make prediction. Van Laarhoven et al[18] introduce Gaussian interaction profile (GIP) to enrich the similarity information then construct a pairwise kernel regression to predict. Gonen et al[19] propose a new Bayesian formulation method which adopts the dimensionality reduction, matrix factorization and binary classification skills. Yuhao et al[20] use the restricted Boltzmann machines learning but this method aims at finding the hidden interaction rather than new interaction. Hailin et al[21] propose a new method using Consistency in Networks which can utilizes large labeled and unlabeled data.

In this paper we propose a label extended semi-supervised learning (LESSL) pridction method which makes prediction from drug view and protein view.We evaluate our method and other existing methods with ten-fold cross-validations on four datasets.To show our method is practicable we use all the data as training set to predict new drug-target interaction.All the results show that our method can achieve better perfromance.

Materials

Chemical Data. Chemical structures of drug compounds come from the DRUG and COMPOUND which are two useful sections in the KEGG LIGAND database. Yamanishi et al[13] go through each drug pairs then use SIMCOMP program to calculate the structural similarity between drug compounds based on the size of common substructures between compounds. The similarity matrix of all drug compounds is denoted S_D here.

Genomic Data. Amino acid sequences of target protein come from the public database KEGG GENES. Yamanishi et al[13] use a normalized version of Smith-Waterman score to calculate the sequence similarity between target proteins. The similarity matrix of targets is denoted as $S_{P.}$

Drug-protein Interaction Data. Yamanishi et al[13] find that the well confirmed drug-target interactions in nuclear receptors, GPCR ,ion channels and enzymes are 90,635,1476,and 2926.The interaction matrix is Y, if e_{ij}=1, the i-th protein and the j-th drug interacts with each other, e_{ij}=0 otherwise. Those known drug-target interactions are regarded as 'gold standard' just like previous studies[13-21].

Label Extended Semi-supervised LearningMethod

Graph-based Semi-supervised Learning.GSSL is a kind of semi-supervised learning method which makes full use of the unlabeled instance.This method goes with this assumption that the change of all the labels should be smooth on the graph.The graph,showing the correlation of each pair of instances,is G=(V,E) where V is the vertex set and E is the edge set.Each vertex in V represents an instance in data set and each edge in E gives the similarity for each datum pair with a non-negative weight.Then GSSL tries to find a function F which satisfies two conditions:(1)the result F is closed to the original labels,and(2)all the labels should change very smoothly on the graph G.The first assumption can be expressed like this:

$$\varphi_1(F) = \frac{1}{2}\sum_{j=1}^{n}\sum_{j=1}^{m}(f_{ij} - y_{ij}) = \frac{1}{2}\|F - Y\| \tag{1}$$

Where Y is the original label matrix. The second assumption can be expressed like this:

$$\varphi_2(F) = \frac{1}{2}\sum_{i,j=1}^{n}\left\|\frac{f_i}{\sqrt{D_{ii}}} - \frac{f_j}{\sqrt{D_{jj}}}\right\|^2 W_{ij} = tr\left(F^T\left(I - D^{-\frac{1}{2}}WD^{-\frac{1}{2}}\right)F\right) = tr(F^TLF) \tag{2}$$

Where F=[f_1,f_2,...f_n],D is a diagonal matrix whose element is $D_{ii} = \sum_{j=1}^{n} W_{ij}$,I is identity matrix,L is Laplace matrix.Then the GSSL model tries to minimize the sum of two parts showed above:

$$\arg min_F(\frac{1}{2}\|F - Y\| + tr(F^TLF)) \tag{3}$$

For drug target prediction problem we follow the assumption that similar drugs always interact with the same targets and similar targets always interact with the same drugs, so in our method we make the prediction from the two point of views: the drug view and the target view.

Protein View Prediction. The matrix Y is very sparse and the number of known targets for each drug changes sharply. Traditional graph-based semi-supervised learning algorithms are extremely dependent on the initial labels. So we can introduce the matrix V which is a node regularizer[22] here .The matrix V=diag(v) comes from the initial label matrix Y:

$$\mathrm{v} = \sum_{j=1}^{d} \frac{Y_{.j} \circ D\vec{1}}{Y_{.j}^{T} D\vec{1}} \tag{4}$$

Where Y.j means the j-th column of Y, the operator ∘means the Hadamard product and $\vec{1}$is a column vector $[1\ldots1]^T$. So we can normalized the interaction matrix Y by V*Y, then we have a normalized label matrix Z=V*Y.

The matrix Z is incomplete with many false negatives, because the interaction information from the well-known database is updated day by day. To overcome this we can extend the matrix Z by letting Y'=Z*S_d. By this matrix multiplication we may get more complete interaction information matrix Y'[23].

We can get an extra drug similarity matrix D form the interaction matrix Y:

$$D_{ij} = \exp(-\gamma(1 - \cos(Y_{.i} * Y_{.i}))) \tag{5}$$

Whereγis a parameter,$y_{.i}$ is a binary vector which is formed by the i-th column of the matrix Y. Now we use the weighted loss function as the first part in model:

$$\varphi_1(F) = \frac{1}{2} \| M_1 \circ (F - V_1 Y S_d)^T (F - V_1 Y S_d) \| \tag{6}$$

V is a node regularizer, S_d is the drug similarity matrix. M is a new weight matrix which gives the weighted loss for those incomplete interaction information. M is defined as follow:

$$M_{ij} = \begin{cases} 1, Y_{ik} = 1 \\ Y_{i.} D_{.j}, Y_{ik} = 0 \end{cases} \tag{7}$$

We set the second part of the objective function defined as follow:

$$\varphi_2(F) = tr\left(F^T \left(I - D^{-\frac{1}{2}} S_p D^{-\frac{1}{2}}\right) F\right) = tr(F^T L_p F) \tag{8}$$

So our objective function is:

$$min_F \varphi(\mathrm{F}) = \frac{1}{2} \| M_1 \circ (F - V_1 Y S_d)^T (F - V_1 Y S_d) \| + \alpha tr\left(F^T L_p F\right) + \beta \| F^T F \| \tag{9}$$

The third term controls the sparsity of F, because to our knowledge a kind of drug interacts with a few proteins. α and β are two parameters which try to balance the importance of each part. To make it simple, we define a matrix Y'=VY$S_{d.}$

Taking the derivative of Eq.(9) with respect to F, we can get the Eq.(10) as follow:

$$\frac{\partial \varphi(F)}{\partial F} = M \circ (F - Y') + \partial L_d F + \beta I F \tag{10}$$

Eq.(10) can be divided into n parts. For the i-th part it can solved like this:

$$\widetilde{(M_{.k}} + \alpha \mathrm{L} + \beta \mathrm{I}) f_{.k} = j_k \tag{11}$$

$$\widetilde{M_{.k}} = diag(M_{.k}), j_k = M_{.k} \circ Y'_{.k} \tag{12}$$

Eq.(11) can be solved in may ways[24]. The time complexity is related to the non-zero elements in Eq.(12).Because $\widetilde{\mathrm{M}}_{.k}, L$ and I have O(n) non-zero elements. In this paper we chose Conjugate Gradient solver to solve .By solving the equation we can get one prediction result F.

Drug View Prediction. Because of the similarity to protein view prediction, we explain it briefly.

We also need to compute the extra similarity of protein by:

$$P_{ij} = \exp(-\gamma(1 - \cos(Y_{i.} * Y_{j.}))) \tag{13}$$

The weight matrix is:

$$M_{ij} = \begin{cases} 1, Y_{ik} = 1 \\ P_{j.}Y_{.i}, Y_{ik} = 0 \end{cases} \tag{14}$$

To make the way of solving model uniform we make a transform operation to Y and F, letting E=F^{T}. So the drug view prediction model is:

$$\varphi(\mathrm{F}) = \frac{1}{2}\left\|M_2 \circ (E - V_2 Y^T S_p)^T (E - V_2 Y^T S_p)\right\| + \alpha tr(E^T L_d E) + \beta \|E^T E\| \tag{15}$$

By solving it we can get another prediction result E. Then final prediction result is

$$\text{Result} = \frac{F + E^T}{2} \tag{16}$$

Experimental Setup

In order to show the performance of our proposed prediction method, we adopt the 10-fold cross-validation method on the four data sets. The 10-fold cross-validation method consists of two parts:(1)drug prediction view: we split the whole drug compounds into 10 parts of roughly equal size, that is,90% of columns in Y are training set. Then we choose one part from the drugs as test set and the remaining 9 parts as training sets. We repeat this procedure 10 times.(2)target prediction view: we split the whole targets into 10 parts of roughly equal size and 90% of rows in Y are training set.

To measure the performance of prediction, we use the traditional AUC and AUPR .From two kinds of cross-validation we get AUCd, AUPRd ,AUCt and AUPRt. And we set parameters $\gamma = 1, \alpha = 0.01, \beta = 0.001$ empirically in our method and we compare our method with other existing methods. The results are below:

Table 1 AUCd by 5*10-fold cross-validation

AUCd	Nuclear receptor	GPCR	Ion channel	Enzyme
BLM	0.693	0.829	0.770	0.781
LapRLS	0.820	0.845	0.796	0.801
NetLapRLS	0.819	0.834	0.783	0.791
KBMF2K	*0.831*	0.844	0.808	0.783
LESSL	0.827	*0.852*	*0.817*	*0.857*

Table 2 AUPRd by 5*10-fold cross-validation

AUPRd	Nuclear receptor	GPCR	Ion channel	Enzyme
BLM	0.194	0.210	0.167	0.092
LapRLS	*0.482*	0.397	0.366	0.368
NetLapRLS	0.481	0.397	0.343	0.298
KBMF2K	0.450	0.357	0.296	0.253
LESSL	0.480	*0.412*	*0.490*	*0.535*

Table 3 AUCt by 5*10-fold cross-validation

AUCt	Nuclear receptor	GPCR	Ion channel	Enzyme
BLM	0.458	0.627	0.881	0.843
LapRLS	0.563	0.788	0.920	0.914
NetLapRLS	0.561	0.787	0.916	0.909
KBMF2K	*0.756*	0.837	0.924	0.889
LESSL	0.748	*0.845*	*0.941*	*0.927*

Table 4 AUPRt by 5*10-fold cross-validation

AUPRt	Nuclear receptor	GPCR	Ion channel	Enzyme
BLM	0.325	0.367	0.641	0.611
LapRLS	0.432	0.508	0.778	0.792
NetLapRLS	0.433	0.503	0.762	0.787
KBMF2K	0.404	0.412	0.725	0.607
LESSL	*0.463*	*0.537*	*0.782*	*0.810*

Although our method does not achieve best results in each data set, we can find that our method improves the AUC and AUPR on most data sets especially those large data sets.

Thenwe use the the whole data set as a training set to predict new interaction.We rank the prediction result by each drug.Taking drug D00348 as an examplewe rank all the targets to drug D00690 by the predicted score and find out those targets with high score.Thenwe query the latest database to check whether those targets with high score actually interact with drug D00348.We list those well predicted drug-target pairs below.

Table 5 The newly confirmed interaction predicted by our method.

Target ID	Drug ID	Rank	Public database
hsa:5916	D00348	3	ChEMBL
hsa:6258	D00348	2	ChEMBL
hsa:2908	D00690	1	KEGG
hsa:2100	D00067	3	KEGG
hsa:2100	D00312	2	KEGG
hsa:2099	D00182	3	ChEMBL
hsa:1129	D01103	3	KEGG
hsa:4985	D00837	1	DrugBank

Table 6 The newly confirmed interaction predicted by our method.

Target ID	Drug ID	Rank	Public database
hsa:5742	D00569	2	DrugBank
hsa:5742	D00448	3	KEGG
hsa:1575	D00542	4	DrugBank
hsa:4128	D05458	3	DrugBank
hsa:1636	D00035	1	SuperTarget
hsa:6331	D00552	5	KEGG
hsa:6328	D05077	2	KEGG
hsa:779	D00438	2	KEGG
hsa:6323	D00512	3	DrugBank

Summary

In this paper we proposed a label extended semi-supervised learning method for predicting new drug-target interaction. The experimental results show that our method can raise AUC average by 0.03 and raise AUPR average by 0.04. At last we use the whole dataset as a training set and predict over 10 new drug-target interactions.To conclude our method is efficient and practicable. We also find that the similarity information is key to prediction. The original way of measuring similarity of drugs can’t capture all the characteristics of drugs. So in the future study we will try to find a more suitable way to measure the similarity between drugs.

References

[1]Hopkins AL. Drug discovery: predicting promiscuity. Nature (2009); 46 2:167–8.

[2] Moriaud F, Richard SB, Adcock SA, et al. Identify drug repurposing candidates by mining the Protein Data Bank. Brief Bioinform(2011); 12(4):336–40.

[3] Dudley JT, Deshpande T, Butte AJ. Exploiting drug-disease relationships for computational drug repositioning. Bioinform(2011);12(4):303–11.

[4]Hopkins,A.L. and Groom,C.R. The druggable genome. Nat. Rev. Drug Discov (2002);1, 727－730.

[5] Kuruvilla FG, Shamji AF, Sternson SM, Hergenrother PJ, Schreiber SL Dissecting glucose signalling with diversity-oriented synthesis and small-molecule microarrays. (2002)Nature 416: 653 – 657.

[6] Whitebread S, Hamon J, Bojanic D, et al. Keynote review:in vitro safety pharmacology profiling: an essential tool for successful drug development. DrugDiscovToday (2005):10(21):1421 – 33.

[7] Zhu S, Okuno Y, Mamitsuka H (2005) A probabilistic model for mining implicit chemical compoundgene relations from literature. Bioinformatics, 21 (Suppl 2):ii245 – ii251.

[8] Klabunde T, Hessler G. Drug design strategies for targeting G-protein-coupled receptors. Chembiochem (2002);3(10):928 – 44.

[9] Sarwar B, Karypis G, Konstan J, RiedlJ.Item-Based Collaborative Filtering Recommendation Algorithms. In: Proceedings of the World Wide Web Conference(2001). pp 285 – 295.

[10] Herlocker JL, Konstan JA, Terveen K, Riedl JT . Evaluating collaborative filtering recommend systems.ACM T Inform Syst(2004) 22: 5 – 53.

[11] Cheng F, Liu C, Jiang J, Lu W, Li W, et al. Prediction of Drug-Target Interactions and Drug Repositioning via Network-Based Inference. PLoSComputBiol(2012) 8(5): e1002503.

[12]Chen,X.et al. Drug-target interaction prediction by random walk on the heterogeneous network. Mol.Biosyst(2012). 6, 1970 – 1978 .

[13] Bleakley K, Yamanishi Y . Supervised prediction of drug-target interactions using bipartite local models. Bioinformatics(2009) 25: 2397 – 2403.

[14] Yamanishi Y, Kotera M, Kanehisa M, Goto S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics(2010) 26: i246 – 254.

[15] Jian-Ping Mei.et al. Drug – target interaction prediction by learning from local information and neighbors.Bioinformatics(2013) 29:238-245.

[16] Masataka Takarabe1.et al. Drug target prediction using adverse event reportsystems: a pharmacogenomic approach.Bioinformatics(2012)28:i611-i618.

[17] Xia Z, Wu LY, Zhou X, et al. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces.BMCSystBiol (2010)4(Suppl 2):S6.

[18] Van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug-target interaction. Bioinformatics (2011)27(21):3036 – 43.

[19] Gonen M Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. Bioinformatics(2012) 28: 2304 – 2310.

[20] Yuhao Wang and Jianyang Zeng. Predicting drug-target interactions using restricted Boltzmann machines.Bioinformatics(2013) 29: i126 – i134.

[21] Hailinchen et al.A Semi-Supervised Method for Drug-Target Interaction Prediction with Consistency in Networks.PLoS ONE (2013).07. 8(5): e62975.

[22] Jun Wang.et al. Graph Transduction via Alternating Minimization.Proceedings of the 25th international conference on Machine learning(2008).Page 1144-1151.

[23] Xiaofei Zhang and Daoaing Dai. A Framework for Incorporating Functional Interrelationships into Protein Function Prediction Algorithms.IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) (2012).Page 740-753.

[24] J.Nocedal and S. Wright, Numerical Optimization. Springer Verlag, (1999).