Local Imbalance based Ensemble for Predicting Interactions between Novel Drugs and Targets

Bin Liu[1], Konstantinos Pliakos[2,3], Celine Vens[2,3], and Grigorios Tsoumakas[1]

[1] School of Informatics, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece
[2] KU Leuven, Campus KULAK, Faculty of Medicine, Kortrijk, Belgium
[3] ITEC, imec research group at KU Leuven

Abstract. Computational prediction of drug-target interactions (DTI) reduces the number of candidate drugs to be verified by the tedious and costly experimental approach and expedites the drug discovery process. The most challenging task for computational DTI prediction methods is to predict interactions between new drugs and new targets due to the unavailability of interacting information for both new drugs and new targets. Although there are several methods that could predict interactions in new drug-target pairs, the accuracy of their predicting results is not adequate. To improve the performance of existing approaches, we propose three ensemble DTI prediction strategies that could accompany any DTI prediction method. The proposed ensemble approaches consist of several DTI prediction models learned on training subsets which have been defined by different sampling strategies. Experiments were conducted on four benchmark datasets and the obtained results indicate that the local imbalance-aware sampling strategy is the most effective.

Keywords: Drug-Target Interaction Prediction · Ensemble Method · Local Imbalance

1 Introduction

Identifying drug-target interactions (DTI) is a key step for the drug discovery process [1]. However, the identification of DTIs via *in vitro* experiments is still costly and time-consuming. Computational approaches, which predict interactions between drugs and targets efficiently, shrink the number of candidate pharmaceuticals for further experimental validation and accelerate the drug discovery procedure. Chemogenomic approaches are the most widely used computational methods. They have attracted extensive interest, because they utilize information from both the drug and the target space. Chemical structure-based compound similarities and protein sequence-based target similarities are the most common information sources that are employed in DTI prediction tasks, mainly due to their effectiveness and availability [1].

DTI prediction is associated with four kinds of prediction settings: providing predictions for pairs of training (known) drugs and targets (S1), new drugs and

training targets (S2), training drugs and new targets (S3), and new drugs and new targets (S4). S4 is related to zero-shot learning [8, 7] and is substantially more challenging than the other three settings [5] because we have no information about the interaction profiles of either the drug or the target of the new pair. Although several DTI prediction methods [3, 6, 5, 4] are able to handle the S4 setting, their results show that there is clearly still room for improvement.

Ensemble methods integrate multiple models and are particularly effective in improving the overall performance and robustness [10]. Here, we propose three ensemble methods that can be coupled with any DTI prediction approach. We specifically aim at improving the accuracy of DTI prediction methods for the S4 setting. The three ensemble methods follow the same framework that aggregates multiple DTI prediction models built upon diverse training sets comprising a subset of drugs and targets, but employ different sampling strategies to choose drugs and targets. Based on their sampling strategy, the three proposed ensemble methods are called *Ensemble with Random Sampling* (ERS), *Ensemble with Global imbalance based Sampling* (EGS), and *Ensemble with Local imbalance based Sampling* (ELS), respectively. Experimental results on four benchmark datasets using two base models show that ELS outperforms the other two ensemble methods as well as the base models.

2 Our Approaches

Firstly, we define the formulation of the DTI prediction problem. Let $D = \{d_i\}_{i=1}^{n}$ and $T = \{t_j\}_{j=1}^{m}$ be the training drug set and target set respectively, where n is the number of training drugs and m is the number of training targets. $\boldsymbol{S}^d \in \mathbb{R}^{n\times n}$ and $\boldsymbol{S}^t \in \mathbb{R}^{m\times m}$ denote the chemical structure-based drug similarity matrix and protein sequence-based target similarity matrix, respectively. $\boldsymbol{Y} \in \{0,1\}^{n\times m}$ is the interaction matrix showing which drugs and targets interact. The DTI prediction method is built on the training set (D, T, $\boldsymbol{S}^d$, $\boldsymbol{S}^t$, $\boldsymbol{Y}$). Given a drug d_u and a target t_v, along with the similarity vector $\boldsymbol{s}_u^d \in \mathbb{R}^n$ and $\boldsymbol{s}_v^t \in \mathbb{R}^m$ indicating similarities between d_u and D and between t_v and T respectively, a DTI prediction method estimates whether d_u and t_j interact or not. More specifically, in this paper we focus on S4 where $d_u \notin D$ and $t_v \notin T$.

The three proposed ensemble methods follow the same framework that can be applied to any DTI prediction method. In the training phase, the sampling probabilities for drugs and targets, denoted as $\boldsymbol{p}^d \in \mathbb{R}^n$ and $\boldsymbol{p}^t \in \mathbb{R}^m$, are initially computed, where $\sum_{i=1}^{n} p_i^d = 1$ and $\sum_{j=1}^{m} p_j^t = 1$. The calculation of the sampling probabilities is different in each of the three methods and will be illustrated later. Then, q base models are trained iteratively. For the i-th base model, the nR sized drug subset D_i is sampled from D without replacement according to $\boldsymbol{p}^d$, i.e. drugs with larger sampling probability have a greater chance to be added to D_i, where R is a user-specified sampling ratio controlling the number of selected drugs. In a similar way, the mR sized target subset T_i is derived from T based on $\boldsymbol{p}^t$. Then, we form a training subset that consists of the similarity sub-matrices and interaction sub-matrix specified by D_i and T_i, upon which the base model M_i is

built. At inference time, given a test drug-target pair (d_u, t_v), every base model gives a prediction, and the final prediction is the average of the outputs from all base models. As the base models are trained based on a subset of drugs and targets, the similarity vector for d_u and t_v is projected to the low dimensional space characterized by the drug and target subset used in the corresponding base model. For example, given a similarity vector $[1, 0.6, 0.8, 0.9, 0]$, its projection on drug subset $\{d_1, d_2, d_4\}$ is $[1, 0.6, 0.9]$.

We now illustrate the calculation of sampling probabilities in our three ensemble methods. ERS employs a sampling probability following the uniform distribution, i.e. $p_i^d = 1/n$ and $p_j^t = 1/m$, where $i = 1, 2, ...n$ and $j = 1, 2, ...m$. This way each drug and target has an equal chance to be selected.

In DTI data, there is typically a small number of interacting drug-target pairs, which is much lower than the number of non-interacting ones, resulting in an imbalanced distribution within the global interaction matrix. To relieve this global imbalance, EGS forms training subsets by biasing the sampling process to include drugs and targets having more interactions. In EGS, the sampling probability of each drug (target) is proportional to the number of its interactions:

$$\begin{aligned} p_i^d &= \frac{\sigma + \sum_{j=1}^{m} Y_{ij}}{n\sigma + \sum_{h=1}^{n}\sum_{j=1}^{m} Y_{hj}}, \quad i = 1, 2, ...n \\ p_j^t &= \frac{\sigma + \sum_{i=1}^{n} Y_{ij}}{m\sigma + \sum_{i=1}^{n}\sum_{h=1}^{m} Y_{ih}}, \quad j = 1, 2, ...m \end{aligned} \tag{1}$$

where σ is a smoothing parameter. By using Eq.(1), the drugs and targets with more interactions are more likely to be selected in the sampling procedure.

Apart from the global imbalance, the degree of imbalance in the local area around a drug (target) [2] could also be used to assess the importance of that drug (target). The local imbalance of a drug d_i for target t_j could be measured as the proportion of $\mathcal{N}_{d_i}^k$ having the opposite interactivity to t_j compared to d_i:

$$C_{ij}^d = \frac{1}{k} \sum_{d_h \in \mathcal{N}_{d_i}^k} [\![Y_{hj} \neq Y_{ij}]\!] \tag{2}$$

where $\mathcal{N}_{d_i}^k$ is the set with the k nearest neighbours of d_i, retrieved by choosing the training drugs having the k largest values in $\boldsymbol{s}_i^d$. Higher C_{ij}^d means that d_i is surrounded by drugs that have opposite interactivity to t_j. In such cases, correctly predicting Y_{ij} using drug similarities would be difficult. By accumulating the local imbalance of d_i for all interacting targets, the local imbalance of d_i, which also indicates the difficulty of d_i, is computed as:

$$LI_i^d = \sum_{j=1}^{m} C_{ij}^d [\![Y_{ij} = 1]\!] \tag{3}$$

Similarly, the local imbalance of t_j is defined as:

$$LI_j^t = \sum_{i=1}^{n} C_{ij}^t [\![Y_{ij} = 1]\!] \tag{4}$$

The local imbalance of a drug (target) reflects its difficulty. The key idea in ELS is that it encourages more *difficult* drugs and targets to be learned by more base models, reducing thereby the corresponding error. In ELS, the sampling probability is proportional to the corresponding local imbalance:

$$\begin{aligned} p_i^d &= \frac{\sigma + LI_i^d}{n\sigma + \sum_{h=1}^{n} LI_h^d}, \quad i = 1, 2, ...n \\ p_j^t &= \frac{\sigma + LI_j^t}{m\sigma + \sum_{h=1}^{m} LI_h^t}, \quad j = 1, 2, ...m \end{aligned} \tag{5}$$

3 Empirical Results

In the experiments, four benchmark DTI datasets, namely Nuclear Receptors (NR), Ion Channel (IC), G-protein coupled receptors (GPCR), and Enzyme (E) [9] are used. To evaluate the predictive performance for interactions between new drugs and new targets, the 2 repetitions of 3×3-fold block-wise CV under S4 [4] is applied and the Area Under the Precision-Recall curve (AUPR) is employed as the evaluation metric. For all proposed methods, R and q are set to 0.9 and 5 respectively. For EGS, $\sigma = 0.1$. For ELS, $k = 5$ and $\sigma = 0.1/k$. The nearest neighbour method MLkNNSC [6] and the matrix factorization approach NRLMF [3] are used as base models with default settings. For MLkNNSC, $k = 3$. For NRLMF, $\alpha = \beta = \lambda_t = \lambda_d = 0.25$, $\gamma = 1.0$, and $r = 50$.

The obtained results in terms of AUPR are displayed in Table 1, where the best methods are highlighted in boldface, and the ones worse than the base model are underlined. ELS is the best method in most cases and able to improve the accuracy of the base models. This is because ELS emphasizes on *difficult* drugs and targets. EGS comes next but is inferior to the base model in 3 cases, indicating that choosing drugs and targets with more interactions to reduce the global imbalance level only works for limited cases. ERS using a totally random sampling strategy is the worst ensemble method.

Table 1. AUPR Results of ensemble methods along with their embedded base models

Base Model	Dataset	Base	ERS	EGS	ELS
MLkNNSC	NR	0.1111(4)	0.1183(3)	0.1243(2)	**0.1273(1)**
	IC	0.1559(4)	0.1594(3)	**0.1663(1)**	0.1596(2)
	GPCR	0.0933(4)	**0.1026(1)**	0.1004(3)	0.1015(2)
	E	0.1517(2)	0.1384(4)	0.1399(3)	**0.1528(1)**
NRLMF	NR	0.1544(3)	0.147(4)	**0.1801(1)**	0.1751(2)
	IC	0.2209(2)	0.219(3)	0.2166(4)	**0.2212(1)**
	GPCR	0.1358(2)	0.1246(4)	0.1348(3)	**0.1362(1)**
	E	0.1948(4)	0.1985(3)	0.2036(2)	**0.204(1)**
Average Rank		3.13	3.13	2.38	**1.38**

Acknowledgements Bin Liu is supported from the China Scholarship Council (CSC) under the Grant CSC No.201708500095.

References

1. Ezzat, A., Wu, M., Li, X.L., Kwoh, C.K.: Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey. Briefings in Bioinformatics **20**(4), 1337–1357 (2018). https://doi.org/10.1093/bib/bby002
2. Liu, B., Tsoumakas, G.: Synthetic Oversampling of Multi-Label Data based on Local Label Distribution. In: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 19). Würzburg (2019)
3. Liu, Y., Wu, M., Miao, C., Zhao, P., Li, X.L.: Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. PLoS Computational Biology **12**(2) (2016). https://doi.org/10.1371/journal.pcbi.1004760
4. Pliakos, K., Vens, C.: Drug-target interaction prediction with tree-ensemble learning and output space reconstruction. BMC Bioinformatics **21**(1), 1V (2020). https://doi.org/10.1186/s12859-020-3379-z
5. Shi, J.Y., Li, J.X., Chen, B.L., Zhang, Y.: Inferring interactions between novel drugs and novel targets via instance-neighborhood-based models. Current Protein & Peptide Science **19**(5), 488–497 (2018). https://doi.org/10.2174/1389203718666161108093907
6. Shi, J.Y., Yiu, S.M., Li, Y., Leung, H.C., Chin, F.Y.: Predicting drug-target interaction for new drugs using enhanced similarity measures and super-target clustering. Methods **83**, 98–104 (2015). https://doi.org/10.1016/j.ymeth.2015.04.036
7. Waegeman, W., Dembczyński, K., Hüllermeier, E.: Multi-target prediction: a unifying view on problems and methods. Data Mining and Knowledge Discovery **33**(2), 293–324 (2019). https://doi.org/10.1007/s10618-018-0595-5
8. Wang, W., Zheng, V.W., Yu, H., Miao, C.: A survey of zero-shot learning: Settings, methods, and applications. ACM Transactions on Intelligent Systems and Technology **10**(2), 1–37 (2019). https://doi.org/10.1145/3293318
9. Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., Kanehisa, M.: Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics **24**(13) (2008). https://doi.org/10.1093/bioinformatics/btn162
10. Zhou, Z.H.: Ensemble methods: Foundations and algorithms. Chapman & Hall/CRC (2012). https://doi.org/10.1201/b12207