Drug-Target Interaction Prediction with Hubness-aware Machine Learning

Krisztian Buza
Brain Imaging Center, Research Center for Natural Sciences
Hungarian Academy of Sciences, Budapest, Hungary
Email: buza@biointelligence.hu

***Abstract*—Prediction of interactions between drugs and pharmacological targets is an important task for which various machine learning techniques have been applied recently. Although hubness-aware machine learning techniques are among the most promising recently developed approaches, they have not been used for the prediction of drug-target interactions before. In this paper, we extend the Bipartite Local Model (BLM), one of the most prominent approaches for drug-target interaction prediction. In particular, we propose to use a hubness-aware regression technique, ECkNN, as local model. Furthermore, we propose to represent drugs and targets in the similarity space. In order to assist reproducibility of our work as well as comparison to published results, we perform experiments on widely used publicly available real-world drug-target interaction datasets. The results show that our approach is competitive and, in many cases, superior to state-of-the-art drug-target prediction techniques.**

I. Introduction

Despite the sophisticated protocols and policies for drug design, development and follow-up, our knowledge about the interactions between drugs and pharmaceutical targets is incomplete in many cases, especially in case of drugs affecting the central nervous system. Knowing more about drug-target interactions will not only contribute to better understanding of the mechanisms how drugs effect, but it is also relevant for adverse effect prediction and repositioning of known drugs, i.e., use of an existing medicine to treat a disease that has not been treated with that drug yet.

The aforementioned fact that many drug-target interactions are unknown for drugs used to threat the disorders of the central nervous system (CNS) underlines the importance of the drug-target prediction problem. On the one hand, CNS plays an essential role; on the other hand, the costs associated with disorders affecting CNS are enormous: solely in Europe, the total annual costs associated with brain disorders is estimated to be approximately 800 billion EUR [1].

As the biochemical validation of hypothesized drug-target interactions is laborious, time-consuming and expensive [2] [3]. Therefore, computational methods have been proposed for the prediction of drug-target interactions [4] [5]. Classic techniques include approaches based on molecular docking [6] [7] [8] and ligand chemistry [9]. A serious limitation of docking-based approaches is that they require information about the three-dimensional structure of candidate drugs and targets which is often not available, especially for G-protein coupled receptors (GPCRs) and ion channels. Additionally, the performance of ligand-based approaches decrease in case if only few ligands are known.

In response to the above limitations of classic approaches, machine learning techniques have been proposed for the prediction of drug-target interactions [10] [11] [12]. One of the most prominent methods is based on Bipartite Local Models (BLMs) [13].

The presence of hubs, i.e., entities that are connected to surprisingly many other entities in a network, has been observed for various biological, chemical and medical networks, see e.g. [14] [15]. Similar observations can be made for drug-target networks as well, e.g., Fig.1 shows the degree distribution for the drugs and targets in the *Enzyme* drug-target interaction network (we will describe the data in Section IV). As one can see, the distributions have long tails, i.e., there are drugs (and targets, resp.) that are connected with surprisingly many targets (drugs, resp.) compared to "average" drugs (targets, respectively).

In the machine learning community, the presence of hubs has been observed in nearest neighbor graphs, see e.g. [16] [17] [18], and hubness-aware classifiers were developed, see [19] for a survey. More recently, hubness-aware regression techniques, including k-nearest neighbor with error correction (ECkNN), were developed that allow for predictions on a continuous scale [20]. Despite the fact that hubness-aware techniques are among the most promising recent machine learning approaches, to our best knowledge, they have not been applied to the drug-target prediction problem previously.

In this paper, we extend the Bipartite Local Model (BLM). In particular, we propose to use ECkNN as local model with drugs and targets being represented in the similarity space. In order to assist reproducibility of our work as well as comparison to published results, we perform experiments on publicly available real-world drug-target interaction datasets. The results show that our approach is competitive and, in many cases, superior to state-of-the-art drug-target prediction techniques.

The rest of this paper is organized as follows: in Section II we review the background necessary to understand our work. In particular, we focus on BLM and ECkNN. Section III presents the proposed approach, followed by its experimental evaluation in Section IV. Finally, conclusions are drawn in Section V.

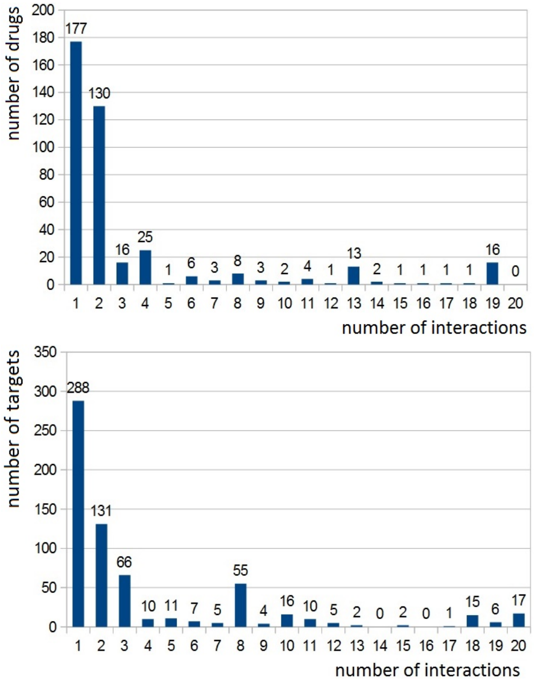

Fig. 1. The degree distribution in the *Enzyme* drug-target network. The horizontal axis corresponds to the number of interactions, whereas the vertical axis corresponds to the number of drugs (in the top) or targets (in the bottom). For example, the first column in the diagram in the tops shows that 177 drugs of the dataset participate in only one of the drug-target interactions. (We note that these drugs do not necessarily participate in the *same* interaction.) In contrast, some drugs (and targets, resp.) participate in surprisingly many interactions, e.g., there are 16 drugs so that each of them interacts with 19 targets.

II. Background

In order to ensure that the paper is self-contained, in this section, the BLM approach for drug-target interaction prediction is reviewed. This is followed by the description of hubness-aware error correction for nearest neighbor regression.

A. BLM: Bipartite Local Model

Bipartite Local Models (BLMs) [13] consider the drug-target interaction prediction problem as a link prediction problem in bipartite graphs. As shown in Fig. 2, the vertices in one of the vertex classes of the bipartite graph correspond to drugs, whereas the vertices in the other vertex class correspond to targets. Each edge e_{ij} of the graph corresponds to a known interaction between drug d_i and target t_j.

When predicting the likelihood of an unknown interaction e_{ij} between drug d_i and target t_j, the model computes two independent predictions that are aggregated subsequently.

The first prediction is based on the relations between d_i and the targets. Each target (except t_j) is labeled as "+1" or "-1" depending on whether or not there is a known interaction between d_i and the target. Then a model is trained to distinguish between "+1"-labeled and "-1"-labeled targets. Finally, this model is applied to predict the likelihood of the unknown interaction e_{ij}.

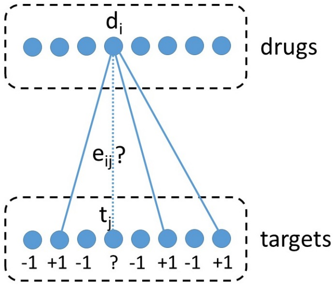

Fig. 2. Bipartite Local Models for the prediction of drug-target interactions.

The second prediction is obtained in a similar fashion, but instead of considering the interactions of drug d_i and labeling the targets, the interactions of target t_j are considered and drugs are labeled. The models that make the first and second predictions are called *local models*.

Finally, the two predictions are aggregated. Originally, Bleakley and Yamanishi took the maximum of the two independent predictions [13], but, in principle, any aggregation function is possible.

We note that various models can be used for each of the independent predictions. While Bleakley and Yamanishi used support vector machines with a domain-specific kernel, we propose to use a hubness-aware regression technique, ECkNN, which is described in the next section.

B. ECkNN: k-Nearest Neighbor Regression with Error Correction

In the last decade, various regression schemes have been introduced, one of the most popular techniques out of them is the k-nearest neighbor regression. When predicting the numeric label on an instance x with k-nearest neighbor regression, the k-nearest neighbors of x (i.e., k most similar instances to x) are determined and the average of their labels is calculated as the predicted label of x. In our case, instances may either correspond to drugs or targets, depending on whether the first or the second prediction of the BLM is calculated.

While being intuitive, k-nearest neighbor regression is well-understood from the point of view of theory, see e.g. [21], [22] and [23] and the references therein for an overview of the most important theoretical results. These theoretical results are also justified by empirical studies: for example, in their recent paper, Stensbo-Smidt et al. found that nearest neighbor regression outperforms model-based prediction of star formation rates [24], while Hu et al. showed that a model based on k-nearest neighbor regression is able to estimate the capacity of lithium-ion batteries [25].

Despite all of the aforementioned advantages of k-nearest neighbor regression, one of its recently explored shortcomings has to be mentioned, namely, the suboptimal performance in the presence of bad hubs. Intuitively, bad hubs are instances that appear as nearest neighbors of many other instances, but have substantially different labels from those instances. For a more detailed discussion, we refer to [20].

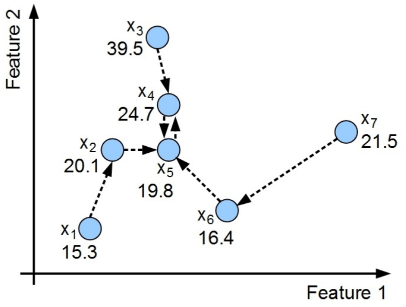

Fig. 3. Example used to illustrate nearest neighbor regression with error correction. The real values next to each instance x_i denote the labels of the instances. In case of drug-target interaction prediction, labels are "+1" (presence of interaction) and "−1" (absence of interaction).

Fortunately, the detrimental effect of bad hubs may be alleviated with relatively simple techniques, such as the error correction technique described below. We define the *corrected label* $y_c(x)$ of a training instance x as

$$y_c(x) = \begin{cases} \frac{1}{|\mathcal{I}_x|} \sum\limits_{x_i \in \mathcal{I}_x} y(x_i) & \text{if } |\mathcal{I}_x| \geq 1 \\ y(x), & \text{otherwise} \end{cases}, \qquad (1)$$

where $\mathcal{I}_x$ denotes the set of training instances that have x as one of their k-nearest neighbors and $y(x)$ is the original (i.e., uncorrected) label of instance x. In k-nearest neighbor regression with error correction (ECkNN), the corrected labels are used instead of the original labels.

Using the example in Fig. 3 we illustrate how the corrected labels are calculated. In Fig. 3 training instances are denoted by circles. They are identified by the symbols $x_1...x_7$. The numeric value next to each instance shows its label. In order to keep the example simple, we use $k = 1$ nearest neighbor to calculate the corrected labels of training instances. In the figure, directed edges point from each instance to its first nearest neighbor. We only present the calculations for x_4 and x_5 as the procedure is the similar in case of the other instances as well. Concretely, the corrected labels of x_4 and x_5 are:

$$y_c(x_4) = \frac{1}{2}(39.5 + 19.8) = 29.65,$$

$$y_c(x_5) = \frac{1}{3}(20.1 + 24.7 + 16.4) = 20.4.$$

A public implementation of ECkNN is available in the PyHubs library.[1]

III. Our approach

Additionally to known drug-target interactions, (i) a domain-specific similarity matrix containing the similarity between drugs and (ii) another similarity matrix containing the similarity between targets are given. This allows to represent drugs (or targets, resp.) in terms of their similarities to all the drugs (or targets, resp.). E.g. the first, second, third... features in the representation of drug d describes the similarity between d and the first, second, third... drug of the dataset. When using this representation, we say that drugs (or targets, resp.) are represented in the similarity space. We propose to represent drugs and targets in the similarity space and use ECkNN as local models in BLM.

Furthermore, in order to obtain the final prediction of BLM, we propose to *average* the two independent predictions of BLM. This is different from the original proposal of Bleakley and Yamanishi who used the *maximum* aggregation function. However, we observed that in case of representing drugs and targets in the similarity space as proposed above, all models, including the baselines, performed slightly better when the *average calculation* was used as aggregation function instead of *maximum*.

TABLE I
Basic properties of the datasets used in the experiments.

Dataset	# Drugs	# Targets	# Interactions
Enzyme	445	664	2926
Ion Channels	210	204	1476
GPCR	223	95	635

IV. Experimental Evaluation

We performed experiments on three publicly available datasets containing drug-target interactions, namely *Enzyme*, *Ion Channel* and *GPCR*.[2] These datasets have been widely used in the literature, see e.g. [10] [11] [12] and [13]. The number of drugs, targets and interactions in these datasets are shown in Tab. I.

In order to assist reproducibility and comparability with published results, we used the same evaluation protocol as Bleakley and Yamanishi [13], i.e., we used leave-one-interaction-out cross-validation and evaluated the predictions in terms of Area Under ROC Curve (AUC) and Area Under Precision-Recall Curve (AUPR).

We implemented the proposed approach in Python. We used the ECkNN implementation from the publicly available PyHubs library and methods from the NumPy machine learning library for the calculation of AUC and AUPR.

As baseline, we used BLM with k-nearest neighbor regression (kNN) *without* error correction as local model. In order to ensure fair comparison between ECkNN and kNN, drugs and targets were represented in the similarity space in both cases. Moreover, we considered $k = 3$ nearest neighbors in both cases. Additionally, we compared our results with the ones published by Bleakley and Yamanishi [13].

Tab. II summarizes our results. Our approach is denoted by ECkNN whereas the baseline is denoted by kNN. As one can see, our approach consistently outperforms the baseline both in terms of AUC and AUPR. Furthermore, compared with the results published by Bleakley and Yamanishi [13], we can conclude that our approach is competitive and, in many cases, superior to the results achieved by the original BLM. For

[1] See also: http://www.biointelligence.hu/pyhubs

[2] See also: http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/.

TABLE II
SUMMARY OF EXPERIMENTAL RESULTS.

Dataset	Method	AUC (in %)	AUPR (in %)
Enzyme	EC*k*NN	95.4	83.7
	*k*NN	94.4	83.5
Ion Channel	EC*k*NN	97.2	85.5
	*k*NN	96.5	82.4
GPCR	EC*k*NN	97.2	62.8
	*k*NN	93.1	61.6

example, on the ion channel dataset, the best model reported in [13] achieved an AUPR of 81.3% whereas our approach has an AUPR of 85.5%. Also the results on the other two datasets outperform most of the results reported in [13].

We note that Bleakley and Yamanishi also report results on the combination of BLM and the kernel regression-based method (KRM). Combination of our approach with other approaches from the literature, such as the aforementioned KRM, is expected to further increase prediction accuracy, however, this is out of scope of the current paper and left for future work.

V. Conclusion and Outlook

In this paper, we considered the drug-target interaction prediction problem which has important applications in understanding the mechanisms of how drugs effect, drug repositioning and prediction of adverse effects. We proposed an extension of BLM, one of the most prominent drug-target prediction models. In particular, we proposed to represent drugs and targets in similarity space and use EW*k*NN, a hubness-aware regression approach as local model in BLM. We performed experiments on widely used publicly available datasets, the results of which show that our approach is competitive and, in many cases, superior to other drug-target prediction approaches from the literature.

On the long term, prediction of drug-target interactions may also play an important role in the realization of the vision of personalized medicine. In order to allow for that, predictions should be made about interactions between drugs and *personalized targets*, e.g. targets related to genetic polymorphisms.

Acknowledgment

This research was performed within the framework of the grant of the Hungarian Scientific Research Fund - OTKA 111710 PD. This paper was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

References

[1] J. Olesen, A. Gustavsson, M. Svensson, H.-U. Wittchen, and B. Jönsson, "The economic cost of brain disorders in europe," *European Journal of Neurology*, vol. 19, no. 1, pp. 155–162, 2012.

[2] S. J. Swamidass, "Mining small-molecule screens to repurpose drugs," *Briefings in bioinformatics*, vol. 12, no. 4, pp. 327–335, 2011.

[3] S. Whitebread, J. Hamon, D. Bojanic, and L. Urban, "Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development," *Drug discovery today*, vol. 10, no. 21, pp. 1421–1433, 2005.

[4] C. J. Manly, S. Louise-May, and J. D. Hammer, "The impact of informatics and computational chemistry on synthesis and screening," *Drug discovery today*, vol. 6, no. 21, pp. 1101–1110, 2001.

[5] B. Bolgár, A. Arany, G. Temesi, B. Balogh, P. Antal, and P. Matyus, "Drug repositioning for treatment of movement disorders: from serendipity to rational discovery strategies," *Current topics in medicinal chemistry*, vol. 13, no. 18, pp. 2337–2363, 2013.

[6] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe, "A fast flexible docking method using an incremental construction algorithm," *Journal of molecular biology*, vol. 261, no. 3, pp. 470–489, 1996.

[7] I. Halperin, B. Ma, H. Wolfson, and R. Nussinov, "Principles of docking: An overview of search algorithms and a guide to scoring functions," *Proteins: Structure, Function, and Bioinformatics*, vol. 47, no. 4, pp. 409–443, 2002.

[8] A. C. Cheng, R. G. Coleman, K. T. Smyth, Q. Cao, P. Soulard, D. R. Caffrey, A. C. Salzberg, and E. S. Huang, "Structure-based maximal affinity model predicts small-molecule druggability," *Nature biotechnology*, vol. 25, no. 1, pp. 71–75, 2007.

[9] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, and B. K. Shoichet, "Relating protein pharmacology by ligand chemistry," *Nature biotechnology*, vol. 25, no. 2, pp. 197–206, 2007.

[10] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa, "Prediction of drug–target interaction networks from the integration of chemical and genomic spaces," *Bioinformatics*, vol. 24, no. 13, pp. i232–i240, 2008.

[11] M. Gönen, "Predicting drug–target interactions from chemical and genomic kernels using bayesian matrix factorization," *Bioinformatics*, vol. 28, no. 18, pp. 2304–2310, 2012.

[12] Z. Xia, L.-Y. Wu, X. Zhou, and S. T. Wong, "Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces," *BMC systems biology*, vol. 4, no. Suppl 2, p. S6, 2010.

[13] K. Bleakley and Y. Yamanishi, "Supervised prediction of drug–target interactions using bipartite local models," *Bioinformatics*, vol. 25, no. 18, pp. 2397–2403, 2009.

[14] A. L. Hopkins, "Network pharmacology: the next paradigm in drug discovery," *Nature chemical biology*, vol. 4, no. 11, pp. 682–690, 2008.

[15] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics*, vol. 12, no. 1, pp. 56–68, 2011.

[16] M. Radovanović, A. Nanopoulos, and M. Ivanović, "Hubs in space: Popular nearest neighbors in high-dimensional data," *The Journal of Machine Learning Research*, vol. 11, pp. 2487–2531, 2010.

[17] N. Tomašev, M. Radovanovic, D. Mladenic, and M. Ivanovic, "A probabilistic approach to nearest-neighbor classification: Naive hubness bayesian knn," in *Proc. CIKM*, 2011.

[18] K. Buza, A. Nanopoulos, and L. Schmidt-Thieme, "Insight: efficient and effective instance selection for time-series classification," in *Advances in Knowledge Discovery and Data Mining*. Springer, 2011, pp. 149–160.

[19] N. Tomašev, K. Buza, K. Marussy, and P. B. Kis, "Hubness-aware classification, instance selection and feature construction: Survey and extensions to time-series," in *Feature selection for data and pattern recognition*. Springer-Verlag, 2015.

[20] K. Buza, A. Nanopoulos, and G. Nagy, "Nearest neighbor regression in the presence of bad hubs," *Knowledge-Based Systems*, vol. 86, pp. 250–260, 2015.

[21] L. Devroye, L. Györfi, A. Krzyzak, and G. Lugosi, "On the strong universal consistency of nearest neighbor regression function estimates," *The Annals of Statistics*, pp. 1371–1385, 1994.

[22] G. Biau, F. Cérou, and A. Guyader, "On the rate of convergence of the bagged nearest neighbor estimate," *The Journal of Machine Learning Research*, vol. 11, pp. 687–712, 2010.

[23] G. Biau, L. Devroye, V. Dujmović, and A. Krzyżak, "An affine invariant k-nearest neighbor regression estimate," *Journal of Multivariate Analysis*, vol. 112, pp. 24–34, 2012.

[24] K. Stensbo-Smidt, C. Igel, A. Zirm, and K. S. Pedersen, "Nearest neighbour regression outperforms model-based prediction of specific star formation rate," in *Big Data, 2013 IEEE International Conference on*. IEEE, 2013, pp. 141–144.

[25] C. Hu, G. Jain, P. Zhang, C. Schmidt, P. Gomadam, and T. Gorka, "Data-driven method based on particle swarm optimization and k-nearest neighbor regression for estimating capacity of lithium-ion battery," *Applied Energy*, vol. 129, pp. 49–55, 2014.