Neighbor Information Integration for DTI Prediction From Heterogeneous Networks

Mert Bayraktar
Department of Computer Engineering Akdeniz University Antalya, Turkey 202151075008@ogr.akdeniz.edu.tr

***Abstract*—Drug-target interaction (DTI) prediction plays a key role in drug development. Many computational approaches for predicting potential DTI based on known DTI have been designed, but their precision needs to be enhanced. Deep learning models offer great performance in DTI prediction. We design a new nonlinear end-to-end learning model that combines diverse data from heterogeneous network data and automatically learns topology-preserving representations of drugs and targets to simplify DTI prediction. We are inspired by recent advances in information passing and aggregation methods that generalize convolution neural networks to mine large-scale graph data and greatly enhance the performance of many network-related prediction tasks.**

***Index Terms*—**

I. Introduction

The drug works by changing the expression of the target protein to achieve the therapeutic effect on the condition. Finding DTI is thus the foundation of drug development. Innovative drug research and development can cost billions of dollars and take more than a decade, and it nearly always fails. As a result, using established DTIs to find prospective drug-target interactions (DTIs) is an important choice for pharmaceutical companies. People are familiar with the qualities of existing drugs, and their safeness is guaranteed. However, biochemical tests to identify novel DTIs have some restrictions in terms of range and throughput. As a result, computational approaches for predicting DTIs have gotten a lot of engagement [1]. In computational-aided drug screening, the three primary groups of prediction methodologies are structure-based, ligand-similarity-based, and machine-learning-based. Structure-based approaches often require three-dimensional protein structures and work poorly for proteins with unknown structures, which regrettably is the situation for the vast majority of targets. To make predictions, ligand-similarity based approaches use common knowledge of known interacting ligands. If the compund of interest is not listed in the library of reference ligands, such approaches will not yield reliable prediction results. Machine learning-based methods that fully leverage latent correlations between relevant properties of drugs and targets have recently emerged as a highly promising strategy for DTI prediction. For instance, to forecast future DTIs, DTI network data has been combined with drug structure and protein sequence data in a network-based machine learning model (e.g., a regularized least-squares framework). Models with improved predictive power have been built in many drug discovery situations (e.g. compound–protein interaction prediction, drug discovery using one-shot learning) as a result of the current increase in deep learning techniques. In this paper, we offer a new framework for predicting new DTIs from heterogeneous data. This model uses non-linear feature extraction using neural networks to combine neighborhood information of the heterogeneous network (HN) built from many data sources through many information passing and aggregation techniques. The extracted feature representations of drugs and targets are then implemented to match the observed networks using a network topology-preserving learning technique. Extensive experiments on multiple challenging and pragmatic DTI prediction scenarios have shown that our end-to-end prediction model outperforms several baseline predictions approaches significantly.

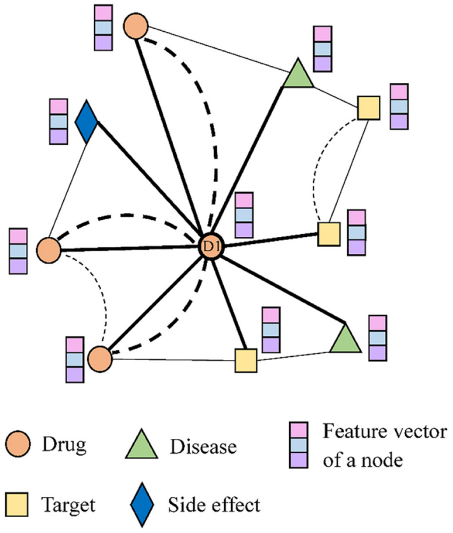

Fig. 1. Example heterogeneous network. Source: Adapted from [1].

II. Related Works

Traditional and deep learning-based approaches have been used in previous DTI prediction research. Traditionally, DTIs are predicted using pharmacological, topological, or semantic similarities based on statistical learning [2]. Using the network structure created by known drug-drug interactions(DDI) and numerous drug classification parameters, Aurel developed predictive pharmacy interaction networks to predict unknown DTIs [3]. The integration of protein-protein interaction (PPI) networks and drug structures, according to Huang, can increase DTI prediction performance [4]. To predict DTIs, Zhang suggested a matrix perturbation technique based on

DDI networks and similar drug properties [5]. To forecast DTIs, Park suggested the random walk with restart approach to simulate signal transmission on protein networks. To forecast DTIs, Park suggested the random walk with restart approach to simulate signal transmission on protein networks. From many perspectives, Zheng estimated and built a framework for drug similarity integration, and he provided a method for selecting positively trustworthy negative samples to forecast DTIs [6]. These methods do not rely on standard methods for forecasting DTI, which are deep learning methods with significant learning representation ability. Kipf researched the convolution operation on graphs based on convolutional neural networks on images as deep learning progressed. He proposed a spectral convolution-based graph convolutional neural network (GCN) method [7]. The approach maintains the network structure and the relationship between the Laplace matrix and the Fourier transform using the Laplace feature spectrum, as well as performing a convolution operation using the Laplace feature spectrum. Many research that uses a deep learning method to predict the side effects of medications generally use the graph convolution methodology since the GCN algorithm has obtained outstanding results in graph link prediction task tests. Jure proposed employing GCN for DDI, PPI, and DPI end-to-end learning to predict DTI [8].

III. Method and Dataset

A. Problem Formulation

The proposed model uses a drug and target-related HN, in which drugs, targets, and other items are represented as nodes, and DTIs and other interactions or correlations are represented as edges, to predict unknown DTIs. The definition of an HN is first presented. Each node v in the node set V belongs to an object type from an object type set O, and each edge e in the edge set belongs to a relation type from a relation type set R. Although our present system may be easily extended to a multi-object-type mapping situation, each node only belongs to one object type. Furthermore, all edges are non-negatively weighted and undirected. Also, the same two nodes can be connected by many edges at the same time, for example, two drugs can be connected by a drug-drug interaction edge and a drug structure similarity edge at the same time [9]. The model seeks to automatically learn a network topology-preserving node-level embedding from an HN that can be utilized to substantially simplify DTI prediction.

B. The Workflow

The three primary steps of the model are as follows: (i) neighborhood information aggregation; (ii) node embedding updating; and (iii) node embedding topology-preserving learning. Each node in a particular HN develops a new feature representation by combining its neighboring information with its own characteristics in Steps (i) and (ii). We enforce topology-preserving node embedding in Step (iii), which is useful for collecting the topological properties of individual nodes for accurate DTI prediction. The mathematical formula for these three phases will be introduced next. Neighborhood information aggregation for node v is defined as: Given an HN G, an initial node embedding function

$$f^0 : V \to \mathbb{R}^d \tag{1}$$

that maps each node v to its d-dimensional vector representation $f^0(v)$, and an edge weight mapping function

$$s : E \to \mathbb{R} \tag{2}$$

that maps each edge e to its edge weight s(e).

$$\sum_{e=(u,v,r)\in E} u \in N_r(v), \quad \frac{s(e)}{M_{v,r}} \sigma\left(W_r f^0(u) + b_r\right) \tag{3}$$

$$a_v = \sum_{r\in R} \tag{4}$$

$$N_r(v) = \{u, u \in V, u \neq v, (u, v, r) \in E\} \tag{5}$$

denotes the set of neighboring nodes connected to $v \in V$ edges of $r \in R$, $\sigma()$ denotes a nonlinear activation function over a single-layer neural network parameterized by weights

$$W_r \in \mathbb{R}^{d\times d}$$

and bias term $b_r \in \mathbb{R}^d$ and

$$M_{v,r} = \sum_{u\in N_r(v), e=(u,v,r)} s(e)$$

stands for a normalization term. Given aggregated neighbor information a_v's for all nodes v's, the process of updating the node embedding is defined as:

$$f^1(v) = \frac{\sigma\left(W^1 \operatorname{concat}\left(f^0(v), a_a\right) + b^1\right)}{\left\|\sigma\left(W^1 \operatorname{concat}\left(f^0(v), a_e\right) + b^1\right)\right\|_2}. \tag{6}$$

The new embedding of node $f^1(v)$ can be obtained using a single-layer neural network parameterized by weights

$$W^1 \in \mathbb{R}^{d\times 2d}$$

and bias term $b^1 \in \mathbb{R}^d$ and a nonlinear activation function σ to nonlinearly transform the concatenation of the original embedding $f^0(v)$ and the neighborhood aggregation information a_v, then normalized by its l2 norm.

Given the node embedding f^1, topology-preserving node embedding learning is defined as:

$$\min_{\left\{f^0(u), W^1, b^1, W_r, b_r,\right.} \sum_{r\in R} \sum_{\substack{u,v\in V \\ \mathrm{G} \\ \left.G_r, H_r, | u\in V, v, r\in R\in E\right.}} \left[s(e) - f^1(u)^\top G_r H_r^\top f^1(v)\right]^2 \tag{7}$$

where

$$G_r, H_r \in \mathbb{R}^{d\times k}$$

are edge-type specific projection matrices. As a result, the model avoids the DTI network, as well as other networks, from being arbitrarily factorized in Step (iii), which can act as a valuable regularizer and improve DTI prediction performance.

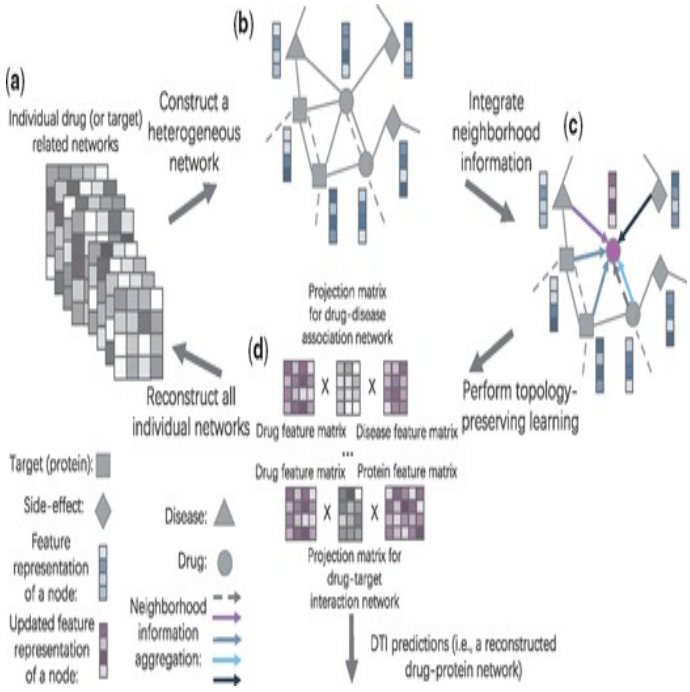

Fig. 2. Workflow of the model. Source: Adapted from: [9].

TABLE I
RESULTS

AUROC	AUPR
0.6982	0.1379

IV. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed model, we utilized 10-fold cross-validation and stratified sampling to verify that the proportion of samples in each category in the training and test sets was the same as in the original dataset. To assess the performance of our approach and baseline methodologies, we used the area under the receiver operator characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). Because it considers both positive and negative data, the receiver operator characteristics (ROC) curve is appropriate for measuring the overall performance of the classifier. In real-world datasets, however, the class imbalance is common. In a DTI network, for example, the number of negative samples is far more than the number of positive ones. The ROC curve in this situation gives an overly optimistic assessment of the effect. Both indicators of the precision-recall (PR) curve, on the other hand, focus on positive samples. People are more concerned about positive samples in class imbalance scenarios, hence the PR curve is often thought to be superior to the ROC curve. Both AUROC and AUPRC are used. The higher the AUROC and AUPRC values, the better the method's performance.

1) Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found at: https://github.com/luoyunan/DTINet

V. CONCLUSION

In this research, we propose a new framework for integrating heterogeneous data from an HN to predict new DTIs. The model uses neural networks to aggregate neighborhood information in the input HN to extract the complex hidden properties of medicines and targets. The model achieves not much higher prediction performance over previous state-of-the-art approaches by optimizing the feature extraction process and the DTI prediction model in an end-to-end way. The effectiveness and robustness of the model have been extensively validated in numerous realistic prediction scenarios, with the conclusion that many of the novel predicted DTIs agree well with earlier literature investigations. Furthermore, the model may easily incorporate other drug and target-related data (e.g., compound–protein binding affinity data).

REFERENCES

[1] Z. X. Liu, Q. F. Chen, et al, "GADTI: Graph autoencoder approach for DTI prediction from heterogeneous network," Frontiers in Genetics, vol. 12, 2021.
[2] T. Liu, J. Cui, H. Zhuang, H. Wang, "Modeling polypharmacy effects with heterogeneous signed graph convolutional networks," Applied Intelligence, vol. 51, pp. 8316-8333, 2021.
[3] Aurel Cami, Shannon Manzi, Alana Arnold, Ben Y. Reis, "Pharmacointeraction Network Models Predict Unknown Drug-Drug Interactions", PLOS ONE, vol.8 (4), Apr. 2013.
[4] Huang, L.C., et al., "Predicting adverse drug reaction profiles by integrating protein interaction networks with drug structures", Proteomics, 2013, 13, (2), pp. 313-324.
[5] W. Zhang, Y. Chen, F. Liu, F. Luo, G. Tian, and X. Li, "Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data," BMC Bioinf., vol. 18, no. 1, Art. no. 18, Jan. 5 2017.
[6] Zheng, Yi, et al. "Inverse similarity and reliable negative samples for drug side-effect prediction," BMC bioinformatics, 2019, 19.13: 554.
[7] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in Proc. ICLR, 2017, pp. 1–14.
[8] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," Bioinformatics, vol. 34, pp. i457–i466, 2018.
[9] F. Wan, L. Hong, A. Xiao, T. Jiang, J. Zeng "NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions," Bioinformatics, vol. 35, pp. 104-111, 2019.