Biomedical Knowledge Graph Embedding with Householder Projection (Student Abstract)

Sensen Zhang[1], Xun Liang*[1], Simin Niu[1], Xuan Zhang[1, 2, 3], Chen Feng[1], Yuefeng Ma[4]

[1] School of Information, Renmin University of China, Beijing, China 100872
[2] Guanghua School of Management, Peking University, Beijing 100871, China
[3] Harvest Fund Management Co., Ltd., Beijing 100020, China
[4] School of Computer, Qufu Normal University, Shandong 276826, China
{sensen0126, xliang, niusimin}@ruc.edu.cn, zhangxuan01@jsfund.cn, 894510507@qq.com, rzmyf1976@163.com

Abstract

Researchers have applied Knowledge Graph Embedding (KGE) techniques with advanced neural network techniques, such as capsule networks, for predicting Drug-Drug Interactions (DDIs) and achieved remarkable results. However, most ignore molecular structure and position features between drug pairs. They cannot model the biomedical field's significant relational mapping properties (RMPs,1-N, N-1, N-N) relation. To solve these problems, we innovatively propose CDHse that consists of two crucial modules: 1) Entity embedding module, we obtain position feature obtained by PubMedBERT and Convolutional Neural Network (CNN), obtain molecular structure feature with Graphic Nuaral Network (GNN), obtain entity embedding feature of drug pairs, and then incorporate these features into one synthetic feature. 2) KGE module, the synthetic feature is Householder projections and then embedded in the complex vector space for training. In this paper, we have selected several advanced models for the DDIs task and performed experiments on three standard BioKG to validate the effectiveness of CDHse.

Introduction

Drug-Drug Interactions (DDIs) occur when patients take multiple medications concurrently and can lead to severe Adverse Drug Reactions (ADRs) that pose a risk to patient safety. With the growing volume of medical literature, numerous DDIs remain undiscovered, resulting in significant implications for medical co-medication practices. Consequently, the identification of DDIs presents a substantial challenge in drug management and development. Accurate prediction of DDIs is crucial for enhancing Biomedical Knowledge Graph (BioKG) and holds significant benefits for patients and public health.

Knowledge Graph (KG)-based methods(Guan, Song, and Liao 2019) are now receiving more attention and achieving the most excellent results. However, the current mainstream KG-based systems ignore the molecular structure and position feature between drug pairs and other feature factors that directly impact drug pair relations. In addition, there are many relation mapping properties (RMPs,1-N, N-1, N-N) problems between drug pairs within BioKG. The existing models, except KG2ECapsule(Su and You 2023), do not consider this aspect, and KG2ECapsule(Su and You 2023)'s sampling method using the Bernoulli distribution makes the model's performance poor and complexity high. To this end, we introduce Householder projections into the study of BioKG for the first time to achieve the rotation of drug pair entities in complex vector space to model the attribute relationships between drug pairs. We also adopt the idea of the neural network to obtain the molecular structure and position features of drug pairs to increase the interpretability of the model.

*Corresponding author

Model Building

Entity Embedding Module

A BioKG is represented as a set of directed triples, ie., BioKG =$\{(h, r, t)\}$. Each triple (h, r, t) consists of a head $h \in \xi$, a relation $r \in \kappa$, and a tail entity $t \in \xi$. We embed the triple into the complex vector space to obtain the embedding vectors $\mathbf{h} \in \mathbb{C}^d$, $\mathbf{r} \in \mathbb{C}^d$, $\mathbf{t} \in \mathbb{C}^d$ of the triple, where $\mathbb{C}$ represents the complex vector space, d represents the dimensionality of the embedding.

Position Feature Firstly, in this paper, we use the WordPiece algorithm(Kudo and Richardson 2018) to process the input drug pair description sentences and obtain the individual words including the drug pair $S = [w_1, w_2, \cdots, w_n]$, where n is the number of words in the drug pair description sentence. Then feed each wordpiece w_i into BERT to get the contextualized embedding $\mathbf{e}_i \in \mathbb{C}^d$, after that we obtain the relative position pair embedding of each word $\mathbf{e}_i^l \in \mathbb{C}^d$, $\mathbf{e}_i^r \in \mathbb{C}^d$, and finally connect the obtained word embedding with the corresponding position pair embedding and feed it into the CNN layer as follows:

$$\mathbf{C}_i = GELU\left(\mathbf{W} \odot \left[\mathbf{e}_{i:i+k-1}; \mathbf{e}^l_{i:i+k-1}; \mathbf{e}^r_{i:i+k-1}\right] + \mathbf{b}\right), \tag{1}$$

where $GELU(\cdot)$ represents the activation function, $\mathbf{W} \in \mathbb{C}^{d^c \times 3d \times k}$ denotes the weight tensor for convolution, k is a window size, $\odot$ is an element-wise product, and $\mathbf{b}$ is a bias term. In addition, in order to transform the output of each filter in the convolution layer to a fixed size vector, this paper uses max-pooling to transform the convolution result as follows:

$$\mathbf{F} = max(\mathbf{C}_i). \tag{2}$$

Model	OGB-Biokg			DrugBank			KEGG		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
BERTKG-DDI	88.35	82.43	86.37	85.24	76.81	84.00	57.73	45.87	47.26
Xin et al.(Jin and Sun 2022)	91.05	84.67	89.26	86.72	86.20	86.29	58.37	45.92	48.27
KG2ECapsule	92.19	89.14	90.64	92.19	89.14	90.64	62.78	47.94	51.31
CDHse	**92.78**	**89.91**	**91.27**	**93.61**	**91.73**	**92.27**	**64.41**	**49.21**	**53.24**

Table 1: The results on datasets. The best results are in bold.

Molecular Structure The molecular graph structures of drugs using GNNs. We represent atoms as nodes and bonds as edges in the drug molecule graph $\mathbf{G}$. We employ the neural molecular GNN method proposed by(Tsubaki and Tomii 2019). The molecular GNN method uses relatively large fragments called r-radius subgraphs or molecular fingerprints to represent atoms with their contexts in the graph. The molecular GNN adopts fingerprint vectors as atom vectors, initializes the vectors randomly, and updates them considering the graph structure of a molecule. The vector of the i-th atom in a drug molecule is a_i, and N_i is the set of its neighboring atoms. The vector a_i is updated in the ℓ-th step as follows:

$$a_i^{\ell} = a_i^{\ell-1} + \sum_{j \in N_i} f\left(\mathbf{W}_b^{\ell-1} b_j^{\ell-1} + \mathbf{b}_b^{\ell-1}\right), \tag{3}$$

where $f(\cdot)$ denotes a ReLU function, and $\mathbf{W}_b$, $\mathbf{b}_b$ are bias term. Then molecular drug carriers are added up with all the atomic carriers, and the resulting carriers are fed into the linear layer as follows:

$$\mathbf{M}_e = f(\mathbf{W}_{out} \sum_{i}^{m} b_i^L + \mathbf{b}_{out}), \tag{4}$$

where m is the number of fingerprints, $\mathbf{W}_{out}$, and $\mathbf{b}_{out}$ are bias term. Then we concatenate different kinds of features as follow:

$$\mathbf{h}^{sf} = \frac{1}{3}(\mathbf{F} + \mathbf{M}_e + \mathbf{h}), \mathbf{t}^{sf} = \frac{1}{3}(\mathbf{F} + \mathbf{M}_e + \mathbf{t}) \tag{5}$$

Knowledge Graph Embedding Module

For each triple (h, r, t), CDHse transforms head entity $\mathbf{h}^{sf}$ and tail entity $\mathbf{t}^{sf}$ with r-specific Householder projections:

$$f(h, r, t) = \left\| \mathbf{P}\left(P_r, T_r\right) \mathbf{h}^{sf} \circ \mathbf{r} - \mathbf{P}\left(P_r, T_r\right) \mathbf{t}^{sf} \right\|, \tag{6}$$

where $\circ$ denotes the Hadamard product, $\mathbf{P}\left(P_r, T_r\right)$ denotes the Householder projections(See appendix for details). The loss function is described as follows:

$$L = -\log \sigma(\gamma - f(h, r, t)) - \frac{1}{k} \sum_{i=1}^{n} \log \sigma(f(h_i^{'}, r, t_i^{'}) - \gamma), \tag{7}$$

where γ is a fixed margin hyper-parameter, σ is the sigmoid function, and $(h_i^{'}, r, t_i^{'})$ is the i-th negative triple.

Experiments

To assess the efficacy of our proposed model, CDHse, we conducted a comprehensive performance evaluation and compared it against state-of-the-art KG-based methods. The empirical results obtained from three datasets are presented in Table 1. Notably, CDHse demonstrates superior performance compared to KG-based models across all three datasets. Specifically, our model exhibits significant improvements over the state-of-the-art KG2ECapsule method when evaluated on the DrugBank and KEGG datasets. Furthermore, our Householder projections approach outperforms the conventional Bernoulli distribution approach.

Conclusions

To enhance the performance of biomedical DDIs prediction, we employ a novel approach that involves embedding drug entity pairs and relationships from BioKG into a complex vector space. Additionally, we incorporate feature information, such as molecular structures and drug descriptions, into the linear combination with the drug entities of BioKG. These combined features, along with the utilization of Householder projections, enable us to effectively model the crucial RMPs within the biomedical domain. Experimental evaluations conducted on various datasets affirm that our model surpasses other state-of-the-art methods in terms of performance.

Acknowledgments

This work was supported by National Natural Science Foundation of China (62072463, 71531012), National Social Science Foundation of China (18ZDA309), Xun Liang is the corresponding author of this paper.

References

Guan, N.; Song, D.; and Liao, L. 2019. Knowledge graph embedding with concepts. *Knowl. Based Syst.*, 164: 38–44.

Jin, X.; and Sun, X. 2022. Extracting Drug-drug Interactions from Biomedical Texts using Knowledge Graph Embeddings and Multi-focal Loss. 884–893. ACM.

Kudo, T.; and Richardson, J. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. 66–71. Association for Computational Linguistics.

Su, X.; and You, Z. 2023. Biomedical Knowledge Graph Embedding With Capsule Network for Multi-Label Drug-Drug Interaction Prediction. *TKDE*, 35(6): 5640–5651.

Tsubaki, M.; and Tomii, K. 2019. Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinform.*, 35(2): 309–318.