RESEARCH

Paper 8 in applied machine learning

Aakash Nepal and Cuong Gia Pham

Full list of author information is available at the end of the article
*Equal contributor

Abstract

Goal of the project: In this project, we were able to get familiar with applying machine learning models to predict relations in knowledge graphs.

Main results of the project: For use-case 1, the results for the multi-class prediction from the models were differing, and for use-case 2, mostly the predicted links made sense by all the models; however, due to the low number of epochs used in our predicting models, we might not have meaningful results, so further research and investigation are needed for reliable results.

Personal key learning :
1 Pham Gia Cuong: new machine learning models, Neo4j, pyKEEN
2 Aakash Nepal: TransE, RotatE, SimplE, Neo4j, pyKEEN

Estimation working hours:
1 Pham Gia Cuong: 10 hours per week
2 Aakash Nepal: 10 hours per week.

Project evaluation: 3.5 out of 5 (due to lack of good computational resources for the project)

Number of words : At least 2000 (without abstract, captions, references)

1 Goal of the project

Knowledge graph completion is a task for the prediction of missing relationships between the nodes in a knowledge graph such as Hetionet. Knowledge graph completion can be useful in predicting links between genes, proteins, illnesses, medications, and other biological processes. Furthermore, we can get insights into drug and target interactions, disease and gene linkages, and other relevant connections by predicting these links. The main goal of this project is to use three different embedding models (TransE, RotatE, and SimplE) from the pyKEEN package and perform multi-class link prediction for the two different use-cases using a subset of the Hetionet dataset. Our first use-case is to see if the methotrexate (MTX) compound in the compound node interacts with genes in the graph that do not presently have any binding links to MTX. MTX is a versatile medicine that is extensively used in the treatment of a variety of disorders, such as cancer, rheumatoid arthritis, psoriasis, inflammatory bowel disease, and viral-mediated arthritis. MTX has a growing interest as a treatment mainly for viral arthritogenic diseases, which are associated with many viruses, including the hepatitis B virus and the human immunodeficiency virus, also called HIV. With the help of multi-class link prediction, our first use-case will help us research and gain more insights into the relationship between MTX and the newly predicted relationships with genes. Our second use-case is to determine whether colon cancer, represented as a Disease node in the knowledge graph, is connected with genes that exist in the network but have no direct connection to colon cancer. Colon cancer is a life-threatening disease that is characterized by the abnormal growth of cells and tissue in the colon. For the better survival of patients, it is necessary to detect colon cancer early. Understanding the roles of critical genes in colon cancer helps unravel the molecular mechanisms for its early development and progression. Using multi-class link prediction, our second use-case will help us gain new predicted links between colon cancer and the genes that might help in further research and other fields.

2 Data

Hetionet is a comprehensive biomedical knowledge network that amalgamates data from 29 diverse databases encompassing genes, compounds, diseases, and more. This integrated network incorporates over 50 years' worth of biomedical information, comprising 47,031 nodes of 11 different types and 2,250,197 relationships of 24 different types. By bringing together this vast array of biomedical data into a unified resource, Hetionet empowers scientists and biologists to create innovative hypotheses, make predictions, and gain valuable insights. Its interconnected nature allows for convenient and holistic exploration of biomedical data across various levels and types, providing a valuable and accessible tool for advancing research and understanding in the field. To use Hetionet dataset for this project, we imported its subset dump file to the neo4j and created a new network.

The data is then queried and organized into a data frame containing three columns: ”source” representing the ID of the compounds in the Compound node, ”target” representing the destination genes connected through the binds link, and ”type” indicating the type of relation. Once these datasets were prepared, we proceeded to embed them using the ”from_labeled_triples” function from the PyKEEN package. This function facilitated the conversion of the labeled triples data into numerical embeddings. Subsequently, the embedded dataset was divided into three sets: the training set, test set, and validation set, with a ratio of 8:1:1. This division allows us to evaluate the performance of the models on unseen data and prevents overfitting during training.

3 Methods

In the methods part, we will use three models named: RotatE, TransE, and SimplE from pyKEEN package. Each of the models will be briefly described. Besides that, the evaluation method is also explained.

3.1 TransE

The TransE model is a popular and foundational model in knowledge graph embeddings topics, introduced in the paper ”Translating Embeddings for Modelling Multi-relational Data” by Antoine Bordes et al.(2013). It aims to represent entities and relations in a knowledge graph as continuous vector embeddings in a low-dimensional space. The main idea behind TransE is to model relations as translations in the embedding space. Equation (1) shows us the mathematical representation of TransE, where h represents the embedding of the head entity (or source), r represents the embedding of the relation, and t represents the embedding of the tail entity. By adding two embedding vectors h and r, we can have a new embedding vector, that is close to the tail entity embedding (or target). During the training, the TransE model learns the embeddings of entities and relations by minimizing the margin-based ranking loss.

$$h + r \approx t \tag{1}$$

3.2 RotatE

Similar to the TransE model, the RotatE model, short for Rotation-based Translational Embedding, is also a knowledge graph embedding model that depicts entities and interactions as complex-valued vector embeddings in a knowledge graph. It introduces rotating patterns to solve earlier models' limitations in properly describing asymmetric interactions. Each relationship is represented as a complex vector with magnitude and angle, and to capture rotational patterns in relationships, a rotational translation is used during scoring. Equation (2) shows us the main idea behind this model. Similar to the above explanation about TransE, h,r, and t respectively represent for head embedding vector, relation embedding vector, and tail embedding vector. The "||||" denotes the L2 norm, where it calculates the Euclidean distance between h*r and t vector. The score of each rotation is calculated by calculation of Euclidean distance between the embedding vector from rotation around the head entity embedding h through relation embedding r and tail embedding vector t. The smaller the distance, the better the model predicts the triple. The model is trained with a margin-based ranking loss function, with the goal of minimizing the distance between genuine triples while increasing the distance between corrupted triples. Despite possible problems such as dealing with large-scale graphs and noisy data, RotatE's ability to capture rich semantics of relationships makes it useful in a variety of knowledge graph-related activities.

$$\text{score}(h, r, t) = ||h \cdot r - t|| \tag{2}$$

3.3 SimplE

The SimplE model is another model for knowledge graph embeddings, short for Simplifying Complex-Valued Embeddings. Like RotatE, SimplE is designed to model knowledge graphs using complex number embeddings, but it aims to address some of the limitations of previous models. Different to RotatE, SimplE represents each relation in the knowledge graph using two separate complex vectors, one for the "left" and one for the "right". With that demonstration, SimplE can capture the symmetric and antisymmetric aspects of relations, which is challenging for models like RotatE which rely on a single complex number of representations for each relation. Equation 3 shows us the mathematical description of SimplE, where h,r, and t represent respectively head, relation, and tail embedding vectors. As mentioned, in the SimplE method, each relation is represented by two complex vectors: "left" and "right" embeddings. Hence, h_l indicates the left embedding vector of the head entity, and t_r indicates the right embedding vector of the tail entity. $Re()$ indicates the scoring function of Simple, where it takes only the real part of the complex number. During training, Simple also learns the embeddings of entities and relations by minimizing a margin-based ranking lost.

$$\text{score}(h, r, t) = Re(h_l * r * t_r) \tag{3}$$

3.4 Prediction and Evaluation

In this project, we want to answer two scientific questions. The first question is " Does MTX compound in the Compound node bind to genes, that are in the graph but do not have any binds link from MTX ?"(use-case 1). To answer this question, we first got the ID of the compound MTX in the Compound node by querying it from neo4j. The prediction process is done by inputting the ID, trained model, training data, and the relation, that you want to predict into the predict_target function from the pyKEEN package. The second question is " Does colon cancer in the Disease node associate with genes, that are in the graph but do not have any associate link from colon cancer ?"(use-case 2). Similar to the first question, the ID of colon cancer is also collected by querying the disease colon cancer from neo4j. The prediction for this question is same as the first question.

The predict_target function calculates the score of the researched node linking to different target nodes. We took the first 5 highest scores and created new relations to target nodes from the researched node. Based on that we can find the new undiscovered relations and their biological meaning.

4 Results and Discussion

After we chose our two use-cases, we simply ran our three models, namely RotatE, TransE, and SimplE on 20 epochs using provided subset of Hetionet dataset. Figure 1 depicts the top 5 results of the RotatE model after multi-class link prediction. Figure 1A shows the predicted new "binds" relationships (edges in yellow) from use-case 1 between the MTX compound (node in blue) and genes (node in yellow). Among the predicted genes, ATP-binding cassette sub-family B member 1 (ABCB1) seems to already have an initial "binds" relationship (edge in pink), which was also predicted by the RotatE model. It was found that the genetic polymorphisms in the gene ABCB1 are closely associated with the dosage of MTX and also influence the treatment response in patients with acute leukemia receiving MTX therapy [1]. One of the genes, namely SLC22A1, is found to be related to another compound called metformin, and SLC22A6, which is one member of SLC22A1's family, is found to be related to the MTX compound[2]. The other genes with predicted links, such as CYP2C9, CYP3A7, and CYP3A4, were not found to be significant for the MTX compound. Figure 1B depicts the multi-class link prediction results for use-case 2 between colon cancer (node in red) and the genes (node in yellow) using the RotatE model. One of the genes that had a new predicted "associates" link (edges in yellow) found by the RotatE model is FAS. It was found that the subpopulation of colon cancer cells has lower FAS expression, which leads to lower sensitivity to FAS ligand-induced apoptosis, which shows a potential role in the behavior of colon cancer stem-like cells [3]. Likewise, the genes such as AKT1, RGS2, BIRC5, and

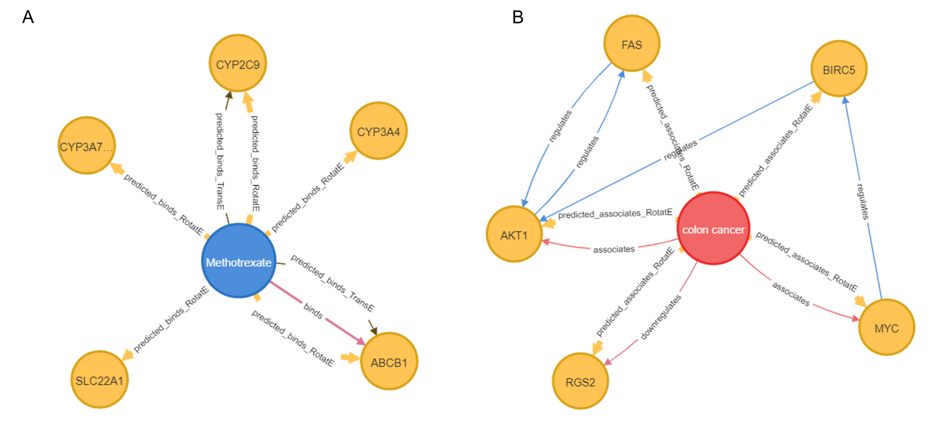

Figure 1 Graphs showing 5 predicted relationships(yellow edges) between respective compound/disease(blue/red node) with the genes(yellow nodes) using RotatE model. A) Graph depicting predicted "binds" relationship by RotatE model for use-case 1. B) Graph depicting predicted "associates" relationship by RotatE model for use-case 2.

MYC, as seen in the graph, that were already related to each other through other relationships such as regulates and downregulates, were also found to be associated with colon cancer[4] [5] [6]. Figure 2 depicts the top 5 results of the TransE model from multi-class link prediction. In Figure 2A, it can be seen that there were 4 new genes with new predicted "binds" relationships, such as CYP3A5, CYP1A2, BCHE, and ADRA2A, and only the gene with the initial "binds" relationship is ABCB1, which was also previously predicted by RotatE, as depicted in Figure 1A. The CYP3A5 gene seems to be associated with the severity of leukocytopenia induced by chemotherapy with MTX in patients with urothelial cancer[7]. However,

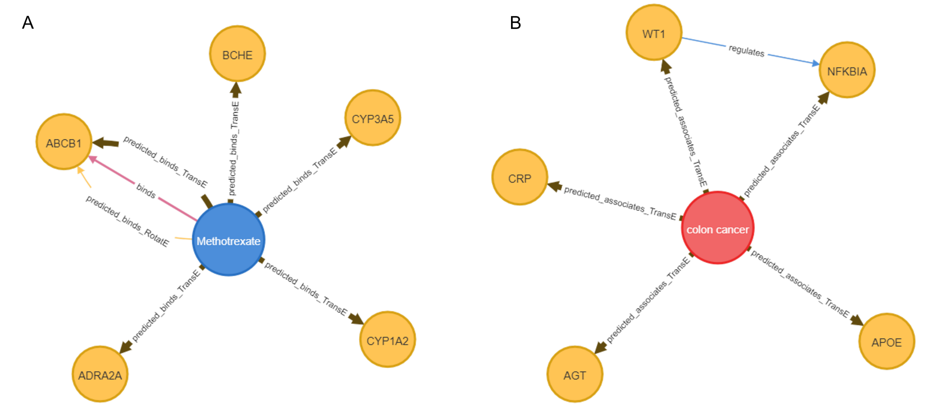

Figure 2 Graphs showing 5 predicted relationships(brown edges) between respective compound/disease(blue/red node) with the genes(yellow nodes) using TransE model. A) Graph depicting predicted "binds" relationship by TransE model for use-case 1. B) Graph depicting predicted "associates" relationship by TransE model for use-case 2.

no association was found between the MTX and the genes BCHE, CYP1A2, and ADRA2A by link prediction done by the TransE model for use-case 1. Figure 2B depicts the top 5 newly predicted "associates" links with their genes, which are WT1, NFKBIA, CRP, AGT, and APOE from the TransE model for use-case 2. Not a single gene among the top 5 predicted links with their gene was found to be the same as the predicted genes found by the RotaE model previously for use-case 2. A study found that the WT1 gene was highly expressed in the majority of tumor tissues compared to normal mucosal tissues. It was also found that the presence of WT1 protein was detected in a significant proportion of colorectal adenocarcinoma cases, which might show a potentially important role of the WT1 gene in the development of colorectal cancer [7]. Other genes connected by predicted relationships by TransE, such as AGT and CRP, except the APOE gene, were also found to be associated with colorectal cancer in use-case 2. Moreover, Figure 3 illustrates the top 5 results of multi-class link prediction done by the SimplE model. In Figure 3A, we can see that genes such as NT5C2, ZNF100, NRAS, CHST3, and CHN1 were predicted by the SimplE model to have a relationship that "binds" with MTX for use-case 1. None of the genes found by the prediction were found to be significantly bound to or associated with the MTX compound. Figure 3B demonstrates the top 5 results found by the SimplE model for use-case 2. Some genes, such as SEL1L and ABHD3, were found to be directly associated with colon cancer, as they are downregulated in colon cancer, as seen in the graph. All other genes, such as COL1A1, TRAP1, and ALDH2B2, were found to be associated with colon cancer.

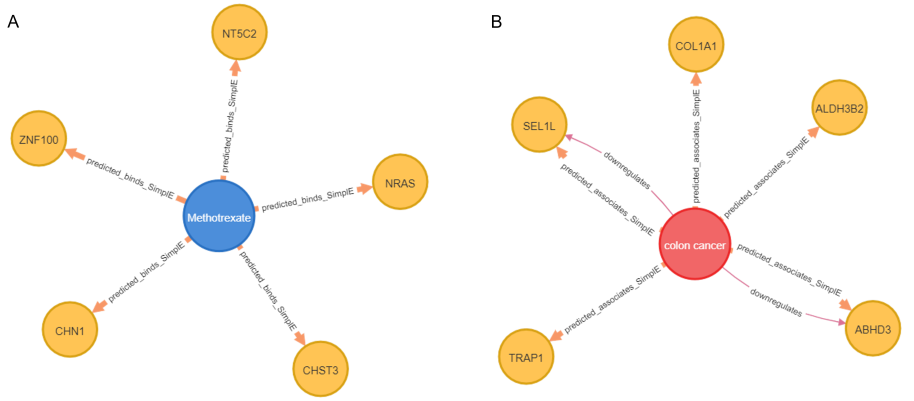

Figure 3 Graphs showing 5 predicted relationships(orange edges) between respective compound/disease(blue/red node) with the genes(yellow nodes) using SimplE model. A) Graph depicting predicted "binds" relationship by SimplE model for use-case 1. B) Graph depicting predicted "associates" relationship by SimplE model for use-case 2.

Overall, according to our findings, for use-case 1, even from the top 5 genes from the predicted links by the TransE and RotaE models, we were able to find one common gene, i.e., ABCB1, which binds the MTX compound. However, 3 out of 5 genes from predicted links by models RotatE and TransE were not found significant, respectively, and none of the genes from predicted links by model SimplE were found significant for use-case 1. For use-case 2, mostly the predicted links make sense as found by all the models; however, due to the low number of epochs used in our predicting models, we might not have meaningful results, so further research and investigation are needed for reliable results.

5 Contributions

1 Pham Gia Cuong: wrote the data, methods part up to RotatE, explained figure 2 in report. Ran TransE and RotatE.

2 Nepal Aakash: wrote the goal of the project, and methods part after SimplE, explained figure 1 and 3, and Ran SimplE and we used the same notebook.

References

[1] Chun-Xia Ma, Yu-Hua Sun, and Hong-Yan Wang. “ABCB1 polymorphisms correlate with susceptibility to adult acute leukemia and response to high-dose methotrexate”. In: *Tumor Biology* 36 (2015), pp. 7599–7606. DOI: 10.1007/s13277-015-3403-5.

[2] Haeun Cheong et al. “Screening of genetic variations of SLC15A2, SLC22A1, SLC22A2 and SLC22A6 genes”. In: *Journal of Human Genetics* 56 (2011), pp. 666–670. DOI: 10.1038/jhg.2011.77.

[3] Wei Xiao et al. “Loss of Fas Expression and Function Is Coupled with Colon Cancer Resistance to Immune Checkpoint Inhibitor Immunotherapy”. In: *Molecular Cancer Research* 17.2 (2019). PMID: 30429213; PMCID: PMC6359951, pp. 420–430. DOI: 10.1158/1541-7786.MCR-18-0455.

[4] Jaclyn F Hechtman et al. “AKT1 E17K in Colorectal Carcinoma Is Associated with BRAF V600E but Not MSI-H Status: A Clinicopathologic Comparison to PIK3CA Helical and Kinase Domain Mutants”. In: *Molecular Cancer Research* 13.6 (2015). PMID: 25714871; PMCID: PMC4978128, pp. 1003–1008. DOI: 10.1158/1541-7786.MCR-15-0062-T.

[5] Zhaohua Jiang et al. “Analysis of RGS2 expression and prognostic significance in stage II and III colorectal cancer”. In: *Bioscience Reports* 30.6 (2010), pp. 383–390. DOI: 10.1042/BSR20090129.

[6] Ming Guo et al. “Identification of the prognostic biomarkers and their correlations with immune infiltration in colorectal cancer through bioinformatics analysis and in vitro experiments”. In: *Heliyon* 9.6 (2023). Published 2023 Jun 12, e17101. DOI: 10.1016/j.heliyon.2023.e17101.

[7] Yusuke Oji et al. “Overexpression of the Wilms’ tumor gene WT1 in colorectal adenocarcinoma”. In: *Cancer Science* 94.8 (2003), pp. 712–717. DOI: 10.1111/j.1349-7006.2003.tb01507.x.