A COMPARISON OF HETEROGENEOUS LINK PREDICTION METHODS ON AN RNA-CENTERED KNOWLEDGE GRAPH

KHUSHBOO CHAUHAN - 14380A
10-12-2024

1.Institution of the Thesis

The Thesis was conducted at the Anacleto Lab in University of Milan, Department of Computer Science, under the supervision of Prof. Marco Mesiti and co-supervision of Dr. Emanuele Cavalleri.

2.Initial Content

A primary challenge addressed in this work is the prediction of the edges within the RNA-KG, a task crucial for uncovering hidden relationships in biological networks. Link prediction, as a method to infer potential new edges between entities, is vital for advancing applications like drug discovery. In this context, the thesis focuses on leveraging heterogeneous graph representation learning (GRL (wei)) techniques to improve edge prediction accuracy.

3.Objectives of the work

The objective of this thesis was to evaluate the performance of various embedding models TransE (Jun), RotatE (z), and RDF2Vec (Ristoski) in classifying node types and predicting edges within the RNA-centered Knowledge Graph (RNA-KG). Specifically, the study aimed to determine how well each model generated embeddings for distinct node types and how effectively these embeddings could be used by classifiers, such as Random Forest and Decision Trees, to predict complex biological relationships and also comparing the results of these model with a graph neural network model GATConv (Roberto). The evaluation was conducted using various performance metrics, including accuracy, precision, recall, and F1-score, to assess the suitability of these models for handling heterogeneous and multi-relational biological data.

4.Description of work performed

The Thesis explores the application of heterogeneous link prediction methods on an RNA-centered knowledge graph (RNA-KG (Cavalleri)), which serves as a resource for understanding RNA molecule interactions within biological systems. RNA-KG is constructed from multiple data sources and incorporates a

variety of biological entities, such as genes, various kinds of RNAs, proteins, diseases, and phenotypes, linked through relationships with precise semantics.

Various methods can be employed to generate node embeddings, which encode the graph structure and biological semantics of RNA-KG. We consider TransE , Rdf2vec, RotatE, as well as Graph Neural Networks (GNNs) based methods. Among them, we considered Graph Attention Networks (GATConv), which are designed to handle the heterogeneous nature of a KG. Traditional machine learning methods, i.e. Decision Trees and Random Forests, are used as classifiers to predict the likelihood of relationships between pairs of nodes based on their embeddings.

5. Technologies involved

The project focused on evaluating machine learning models using RNA-KG as input for node classification and link prediction tasks. Models like TransE, RotatE, RDF2Vec, and GATConv implemented in PyTorch and GRAPE (de) libraries were considered. For classification, Random Forest classifier and Decision Tree were employed to evaluate the embeddings generated by these models. Visualization techniques such as t-SNE and common metrics for classification tasks such as precision, recall, and F1-score were used to analyze and interpret model performance. All experiments were conducted on a T4 GPU support.

6.Skills and results achieved

The results from the node type classification showed that TransE outperformed the other models in terms of accuracy, precision, recall, and F1 scores across all views, particularly for miRNA, Gene, and Disease nodes. RotatE showed moderate performance with decent precision but struggled with recall, especially for miRNA nodes. RDF2Vec faced significant challenges, often producing 0.00 precision and recall due to data sparsity and class imbalance, particularly in View 2. These findings highlight that while TransE is the most reliable model for node type prediction, RDF2Vec needs further improvement to handle the sparsity and imbalance present in the full graph.
GATConv demonstrated superior performance in predicting complex edge types. Random Forest, with its ensemble approach, outperformed Decision Trees in accuracy and precision but required more computational resources. Decision Trees, though faster, faced challenges like overfitting, especially with sparse edge types. TransE and RotatE, while effective for symmetric relationships, struggled with multi-relational data, revealing the limitations of static embeddings in heterogeneous graphs.

One of the significant problems encountered during the node type prediction task was the occurrence of 0.00 precision, recall, and F1 scores, particularly in RDF2Vec and RotatE models when evaluated on the full graph. This issue arose due to data sparsity, class imbalance, and noise in the large, heterogeneous RNA-KG, which led to misclassifications and an inability to correctly predict underrepresented node types like miRNA and Gene. The node classification results across different views show that performance improved when subgraphs with selected node types (miRNA, Gene, and Disease) were used, rather than using the complete graph of a particular VIEW. For instance, the accuracy for View 0 was 0.93 when using TransE embeddings, and 0.82 for RotatE embeddings, highlighting the benefit of focusing on a smaller, more relevant set of nodes. The models performed significantly better on these subgraphs, as they reduced the impact of imbalanced node distributions seen in the full graph.

While most issues were addressed, some challenges, like the overfitting of Decision Trees and the inability of static embeddings to capture the dynamic nature of RNA-KG, remain unsolved. Future improvements could

involve integrating dynamic graph models to account for evolving relationships and further refining attention mechanisms in GATConv to enhance both predictive accuracy and model interpretability.

7. Bibliography

1. Cavalleri, E. et al. “RNA-KG: An ontology-based knowledge graph for representing interactions involving RNA molecules.”
2. de, Lima. et al.“GRAPE: Grammatical Algorithms in Python for Evolution.”
3. Jun, Feng. et al.“Knowledge Graph Embedding by Flexible Translation.” 2016.
4. Ristoski, p.et al. “RDF2Vec: RDF Graph Embeddings for Data Mining.” 2016.
5. Roberto, Corizzo.et al. “Distributed Node Classification with Graph Attention Networks.”
6. wei, ju. et al. “A Comprehensive Survey on Deep Graph Representation Learning.” *A Comprehensive Survey on Deep Graph Representation Learning*, 2024, https://arxiv.org/abs/2304.05055.
7. z, sun. et al. “RotatE for Knowledge Graph Embedding.” 2019.