Unsupervised Link Prediction in Sparse Biomedical Knowledge Graphs for Clinical Insight Enhancement

Isaac Kobby Anni

DATA 7300 Final Project Report | Fall 2024

December 10, 2024

1. Background
2. Dataset
3. Methodology
4. Model Architecture
5. Experiment and Results
6. Conclusion, Recommendations

1 Background

2 Dataset

3 Methodology

4 Model Architecture

5 Experiment and Results

6 Conclusion, Recommendations

Main Focus

- Biomedical knowledge graphs: Tools to analyze relationships between diseases, drugs, symptoms, and genes, [1].

Main Focus

- Biomedical knowledge graphs: Tools to analyze relationships between diseases, drugs, symptoms, and genes, [1].
- Challenge: Graph sparsity limits clinical applications.

Main Focus

- Biomedical knowledge graphs: Tools to analyze relationships between diseases, drugs, symptoms, and genes, [1].
- Challenge: Graph sparsity limits clinical applications.
- Example: Hetionet (47,031 nodes, 2,250,197 edges; density: 0.10%).

Main Focus

- Biomedical knowledge graphs: Tools to analyze relationships between diseases, drugs, symptoms, and genes, [1].
- Challenge: Graph sparsity limits clinical applications.
- Example: Hetionet (47,031 nodes, 2,250,197 edges; density: 0.10%).
- **Objective:** Predict unknown links using unsupervised techniques to enhance clinical insights.

Objectives

- Develop an unsupervised graph-based framework for predicting unknown links in Hetionet.

Objectives

- Develop an unsupervised graph-based framework for predicting unknown links in Hetionet.
- Leverage GraphSAGE and PubBERT for structural and semantic feature learning.

Objectives

- Develop an unsupervised graph-based framework for predicting unknown links in Hetionet.
- Leverage GraphSAGE and PubBERT for structural and semantic feature learning.
- Validate the models effectiveness for clinical applications like drug repurposing and disease diagnosis.

Related Work

- **Hetionet:** Aggregates diverse biomedical data for applications like drug discovery and precision medicine.
- **Traditional Approaches:** Heuristic methods like common neighbors, supervised embeddings (Node2Vec, DeepWalk).
- Graph Neural Networks (GNNs) like GCNs, GATs, and GraphSAGE.
- Addressing sparsity in biomedical graphs with unsupervised methods.

1. Background
2. Dataset
3. Methodology
4. Model Architecture
5. Experiment and Results
6. Conclusion, Recommendations

Dataset Overview

- Source: Hetionet

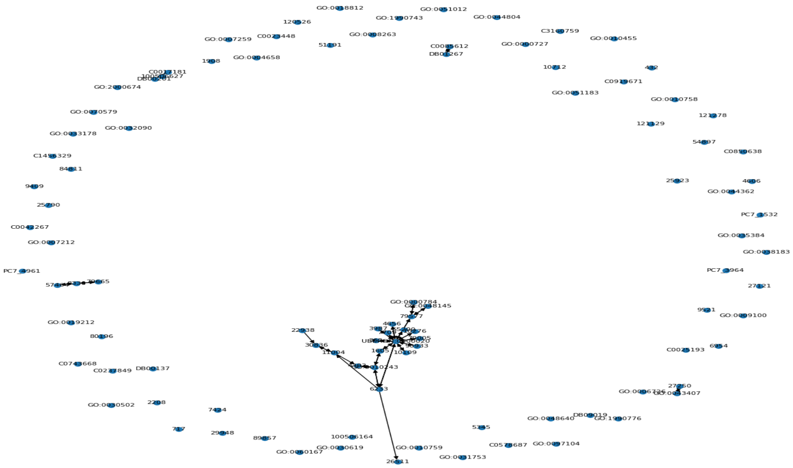
Dataset Overview

- Source: Hetionet
- **Nodes: Diseases, drugs, genes, symptoms, side effects.**

Dataset Overview

- Source: Hetionet
- **Nodes: Diseases, drugs, genes, symptoms, side effects.**
- **Edges: Relationships like treats, interacts, associates.**

Dataset Overview



Dataset Stats

Nodes: 47,031, Edges: 2,250,197, Density: 0.10%.

Preprocessing: Extracted **k-core subgraph** to reduce sparsity.

1. Background
2. Dataset
3. Methodology
4. Model Architecture
5. Experiment and Results
6. Conclusion, Recommendations

Methodology

Data Preprocessing:

- Sampled 60% connected, 40% unconnected nodes.

Graph Representation Learning:

Link Prediction:

Methodology

Data Preprocessing:

- Sampled 60% connected, 40% unconnected nodes.
- Applied k-core decomposition.

Graph Representation Learning:

Link Prediction:

Methodology

Data Preprocessing:

- Sampled 60% connected, 40% unconnected nodes.
- Applied k-core decomposition.

Graph Representation Learning:

- Node features: Encoded with PubMedBERT.

Link Prediction:

Methodology

Data Preprocessing:

- Sampled 60% connected, 40% unconnected nodes.
- Applied k-core decomposition.

Graph Representation Learning:

- Node features: Encoded with PubMedBERT.
- Trained GraphSAGE on k-core subgraph.

Link Prediction:

Methodology

Data Preprocessing:

- Sampled 60% connected, 40% unconnected nodes.
- Applied k-core decomposition.

Graph Representation Learning:

- Node features: Encoded with PubMedBERT.
- Trained GraphSAGE on k-core subgraph.
- Generated embeddings via unsupervised learning.

Link Prediction:

Methodology

Data Preprocessing:

- Sampled 60% connected, 40% unconnected nodes.
- Applied k-core decomposition.

Graph Representation Learning:

- Node features: Encoded with PubMedBERT.
- Trained GraphSAGE on k-core subgraph.
- Generated embeddings via unsupervised learning.

Link Prediction:

- Similarity scoring, thresholding, *validation.*

1. Background
2. Dataset
3. Methodology
4. Model Architecture
5. Experiment and Results
6. Conclusion, Recommendations

Model Architecture

Components:

- PubMedBERT: Captures semantic features.

Layers:

Training:

Model Architecture

Components:

- PubMedBERT: Captures semantic features.
- GraphSAGE: Aggregates neighborhood information.

Layers:

Training:

Model Architecture

Components:

- PubMedBERT: Captures semantic features.
- GraphSAGE: Aggregates neighborhood information.

Layers:

- Three SAGEConv Layers with ReLU activation.

Training:

Model Architecture

Components:

- PubMedBERT: Captures semantic features.
- GraphSAGE: Aggregates neighborhood information.

Layers:

- Three SAGEConv Layers with ReLU activation.

Training:

- Positive and negative sampling.

Model Architecture

Components:

- PubMedBERT: Captures semantic features.
- GraphSAGE: Aggregates neighborhood information.

Layers:

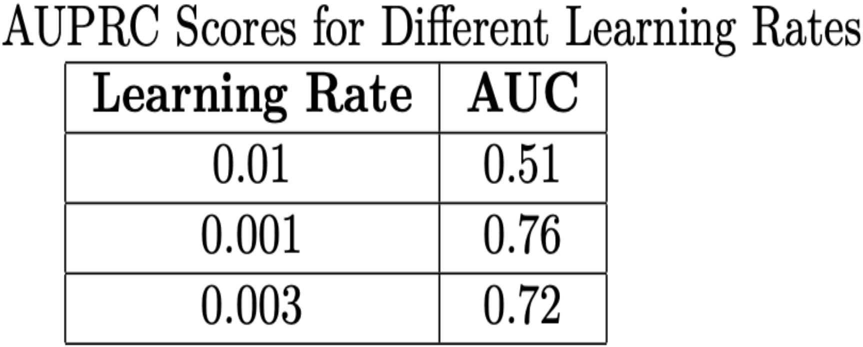
- Three SAGEConv Layers with ReLU activation.

Training:

- Positive and negative sampling.
- Loss function: Binary cross-entropy.

1. Background
2. Dataset
3. Methodology
4. Model Architecture
5. Experiment and Results
6. Conclusion, Recommendations

Experiment & Results



AUPRC Scores for Different Learning Rates

Learning Rate	AUC
0.01	0.51
0.001	0.76
0.003	0.72

Figure 2:

Limitations and Challenges

- Sparsity: Limits link prediction.
- Data Imbalance: Affects negative sampling.
- Performance Limitations: Modest AUPRC (0.508).
- Threshold Sensitivity: High threshold excludes plausible links.

1 Background

2 Dataset

3 Methodology

4 Model Architecture

5 Experiment and Results

6 Conclusion, Recommendations

Conclusion

- Demonstrated utility of GraphSAGE and PubBERT for sparse graphs.
- Challenges highlight areas for future improvement in graph-based machine learning.

Recommendations

- Graph augmentation methods to address sparsity.
- Experiment with sparsity reduction methods.

Future Directions

- Validate predicted links with biomedical experts.
- Experiment with advanced architectures (e.g., GATs).
- Consider other embedding methods for text features of node.

References

1. Himmelstein DS, Lizee A, Hessler C, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife 2017;6:e26726.

Thank you!