Drug Recommendation System using Hetionet Dataset

1. Introduction

In this report, the model aims to develop a drug recommendation system, specifically referring to two tasks: an alternative drug for disease and drugs with side effect constraints. The Hetionet dataset is here for use in one's modelling between the two.

2. Methodology

2.1 Data Set

Data were extracted from the Hetionet dataset, which is a heterogeneous network of biological information. The data were processed to get drug-disease relationships that spell out as triples (head, relation, tail).

2.2 Task 1: Drug Recommendation

Method 1:

Model: TransE, a knowledge graph embedding model.

Process: The model is trained on the Hetionet dataset. It learns the embeddings for the diseases and drugs. For a given disease, it recommends drugs that have the most similar embeddings with that disease.

Evaluation: The best method of performance evaluation is Hits@5, wherein the first five results will be compared with the reference data.

Method 2:

Model: TransE

Process: Similar to Method-1 but does not take into account the drugs having direct connections with disease under the dataset.

Evaluation: NDCG@10 and Hits@10 are used, thereby assessing relevance and ranking quality of the top 10 recommendations.

2.3 Task 2: Drug Recommendations with Side Effect Constraints

Method 1:

Model: Graph Convolutional Network (GCN)

Process: A model is created for GCN, which will predict interaction between drugs and diseases. The predicted side effect constraints will apply afterward, where screened out drugs are above a given similarity threshold.

Evaluation: As it compares recommendations with reference data, the specific metric for this is not mentioned in the provided output.

Method 2:

Model: Relational Graph Convolutional Network (R-GCN)

Process: An R-GCN is used to consider the different kinds of relationships of the Hetionet dataset. Side effect risks are precomputed and then recommendations are done on top of those risks.

Evaluation: The measure used here is Hits@3.

3. Results

3.1 Task 1

Method 1: Sample showing recommendations for the top 5 connections is provided for three diseases. The actual scores reached for Hits@5 are not printed in full but are calculated for the purposes of evaluation.

Method 2: The top 10 alternatives for 2 diseases are listed against similarity scores. The Hits@10 score is 1.0 for disease Id 85, while the NDCG@10 score is 1.0 for disease Id 131, meaning perfect matching is done on both recommendations.

3.2 Task 2

Method 1: Sample recommendations show up for 3 diseases and their top 5 drugs with similarity scores given under constraints of side effects. The evaluation metric and scores are not specified.

Method 2: Top 3 drug recommendations for 3 diseases with relevant scores and candidate IDs are produced, and the evaluation is done using the Hits@3 metric.

4. Conclusion

The project has successfully demonstrated various methodologies applicable to drug recommendation. The comparison indicates that one of these models, TransE, shows a high success rate in recommending alternate drugs.