Link Prediction on EHR data using Biomedical knowledge base

Samskruth Reddy Padigepati, Abhinav Shankaranarayanan Venkataraman
Department of Computer and Information Science and Engineering
University of Florida
Gainesville, Florida 32608,
spadigepati@ufl.edu, abhinavsv3@gmail.com

Introduction

- Electronic health records consists of patient chart information in a digital format. It consists of demographics, vital signs, laboratory tests, medications, physician or nursing notes, discharge summary etc. While the electronic health records are primarily used for storage, retrieval of patient data, the data also has secondary usage in disease research.
- Despite the advantages, challenges in the analysis arise due to the missing data. The missing data may arise due to lack of collection or lack of documentation. Lack of collection arises when a patient is never asked about a condition or never diagnosed with a condition. For example, A person has heart disease but is never diagnosed. Lack of documentation arises when the patient is asked or diagnosed but it is not recorded. This kind of missing data is a huge challenge in analysis since some of the important features are missing.
- In this project we are trying to predict the missing links in the EHR data. We are predicting missing links due to lack of collection and documentation with the help of rule mining on EHR data. With the help of the mined rules we are able to predict the most common diagnosis in the EHR data. We have also integrated a biomedical knowledge base to the EHR data and used the information for predicting the diagnosis which are not that common in the EHR data.

Methods

- We pass the training knowledge graph through the AMIE rule mining system.
- AMIE is a rule mining algorithm that extracts cloea logical rules based on the support in the knowledge base. AMIE iteratively extends rules with the help of mining operators, minimum support threshold, confidence, head search and covers for the rules in the search space. We have executed AMIE with minimum head coverage set to 0.001.
- The rule mining task is executed in 2 parts: 1. Using MIMIC knowledge graph 2. Using UMLS + MIMIC knowledge graph
- For each patient, we have predicted the diagnosis by taking all the rules with rule head having “hasDiagnosis” relationship and whose body is relevant to a patient and then by taking the diagnosis mentioned in the rule head as the prediction. If multiple predictions are done for an instance, then the one with the higher confidence is taken.

Evaluation

- We have evaluated the rules on the subgraph set aside for testing. For the purpose of evaluation, for each patient node in the test subgraph, one of the link to the diagnosis node is masked while keeping links to the other diagnosis same and then the whether there is a link or not between the patient and the masked diagnosis node is to be predicted. This process is repeated for all other links to diagnosis nodes connected to that patient and prediction is done each time.
- The prediction with the help of rules is done as follows. Based on the procedures, prescriptions and all the diagnosis except the one which is masked, relevant rules are filtered from all the rules and then the diagnosis is predicted with the help of the rule head. Even while predicting directly from UMLS relationships, the similar process is done.

Discussion

- It can be seen that the prediction accuracy for all the diagnosis is less when compared to the top 10 and 100 frequent diagnosis. This result is expected because out of all the 6000 diagnosis, only 100 have more than 1000 patients. It can be seen that prediction performance based on rules mined from MIMIC+UMLS knowledge graph has not improved marginally when compared to performance of the prediction based on rules from the MIMIC knowledge graph. But by using the relationships in UMLS directly for prediction the performance is increased.
- In our experiments we have seen that, SNOMED-CT can give us double the number of relationships between the entities than the ICD9 coding system. This motivates us to map all the ICD9 diagnosis and procedures in MIMIC to SNOMED-CT codes to leverage the relationships which otherwise would not have been found in ICD9 mapping.

Conclusions

The experiments conclude that the EHR data can be used for predicting the discharge diagnosis which might be missing in the data due to lack of documentation or lack of collection, by finding patient similarities. Our method uses rule mining to find co-occurrences between patients, diagnosis, treatments and procedures and predicts discharge diagnosis based on the rules. Our experiments also show that by integrating with biomedical knowledge base to the EHR data, new information can be deduced based on information in the EHR.