Building Clinical Decision Support System for Cardiovascular based on Knowledge Graphs

Dianchen Zheng dz2424
Xinyi Wang xw2657
Research Proposal
2/14/2020

Abstract: The purpose of this project is to find the related symptoms and indicator of cardiovascular disease and also its common treatment. The method used in the project will include natural language processing and dynamic time warping. The goal of this project is building a knowledge graph about the Cardiovascular disease based on the patients data and generating future study such as visualizing the diagnosis and treatment outcome.

1. Background (Review of Related Literature):

Cardiovascular disease can refer to a number of conditions: Heart disease, Heart attack, Heart failure, Arrhythmia, ischemic stroke and hemorrhagic stroke. Symptoms are often a precursor to diagnosing a disease. The same is true for heart disease, and patients' feelings can also provide important clues to doctors. However, in real life, it is common to associate certain irrelevant symptoms with heart disease. Many people think that heart palpitations are heart disease. In fact, despite the palpitations, the heart is healthy. Therefore, it is necessary to make clear the symptoms of heart disease in order to arouse people's attention and eliminate unnecessary doubts.

In addition to clinical symptoms and signs, the diagnostic indicators of cardiovascular disease mainly rely on medical testing technology. Although there are echocardiography and magnetic resonance imaging (MRI) in the diagnosis, these tests are expensive and not suitable for dynamic monitoring. Of all the methods, electrocardiogram (ECG) is still the most widely used and inexpensive method, but ECG still has weaknesses in the complicated conditions. With the continuous development of basic medicine and clinical medicine, many cardiovascular disease markers have been applied in the clinic one after another, and have played an important role in the diagnosis of heart disease, risk assessment, efficacy observation, and prognosis estimation.

Knowledge graphs provide complex associations between items. Based on the knowledge graph, the recommendations from the system have several benefits: First one is precision. The recommendation system provides additional links between items, which improves the accuracy of the indicators in the recommendation system. Second is diversity. Using the edges of multiple relations in the knowledge graph to expand the user's interest set outwards can greatly improve the diversity of recommendation results. Third is explainability. The knowledge graph also links the user's historical interests and recommendation results, which provides an additional explanatory source.

2. Introduction to the Project:

First, building a knowledge graph using the information of patients that could reflect the symptoms and indicators related to Cardiovascular disease. For example, Creatine kinase (CK), creatine kinase isoenzyme (CK-MB), etc. are the cardiac injury markers. These markers are normally present in cardiomyocytes and are released into the blood after the occurrence of a myocardial infarction. If elevated levels of these substances are found in the blood, it indicates the presence of myocardial damage. Using the values of several indicators in the data to predict the next medical event, even the next diagnosis based on the patient's previous medical records.

The obstacle of this project is that different diseases may have similar symptoms, which means that the diagnostic indicators from the patient may project to the other disease(which is not true). Since the uncertainty of the medical decision, it would be a challenge to provide accurate and interpretable results.

3. Introduction to the Dataset:

The data mainly comes from the hospitals in TAIPEI and research institutes, and detailed data information has not been obtained yet. Some important information that was already known in the datasets are the patient number, gender and age, time of consultation, symptoms, diagnosis results and treatment. More detailed information is needed, such as the indicators from the EGC, diagnostic indicators about the blood testing, etc. to generate the knowledge graph.

4. Introduction to Algorithms

Based on literature review and our project goals, we can regard this problem as building a recommendation system based on knowledge graph, since we need to predict the patient's potential disease based on his historical medical records and other patients' diseases records. Our group divides this work into five steps.

First, based on related work, we can use NLP methods to extract the disease and symptom entities and their relationships from literature, Wikipedia or by domain knowledge. We are not focused on this part of the job, but we may construct a kg demo to ensure we could run through the whole algorithms pipeline.

Second, we need to embed events based on domain knowledge graph, by which we could represent disease events more accurately and meaningful. Here is an example of event graph embedding, as we can see, disease A and B have more common symptoms, which means if a patient has disease A, he may get disease B either, this Medical phenomenon is known as complication. Therefore, we need to encode similar diseases in similar vectors, so we use the Deep Walk algorithm to encode events.

Third, after embedding events, we need to compare one patient with the other similar patients to do retrieval, which may benefit our decision since similar patients may have similar diseases. However, different patients have different lengths of medical records and we need to neutralize time effect by

dynamic time warping methods. Here is a simple example, the two lines in group 1 and group 6 have a similar trend. However, if they compute their distance directly, the distance is super large. So we need to use dynamic time warping to find a path which could minimize their distance and use this criterion to compute the actual distance. In this way, we could compute the distance of all patients and use clustering methods to find the top-k similar patients.

Fourth, we make event sequence slicing of each patient as input, we could set a time slicing window to train this recurrent neural network and the label is the center event of the patient. The loss function we could choose cross entropy and the output is a vector with the length of the number of diseases. When we get the output vector, we could know the probability of potential disease he may get.

Finally, we need to finish the visualization part, we may use some UI tools to finish this part of work.

4. Plan:

Milestone1: Get the knowledge about the Cardiovascular disease, overview the dataset.
Milestone2: Data Cleaning, Knowledge Graph
Milestone3: Graph Embedding, Dynamic Time Warping
Final: Prediction and Visualization

Reference:

[1] Zhuochen Jin, Jingshun Yang, Shuyuan Cui. CarePre: An Intelligent Clinical Decision Assistance System. arXiv:1811.02218v1
[2] Shunan Guo, Zhuochen Jin, David Gotz, Fan Du, Hongyuan Zha, and Nan Cao. Visual Progression Analysis of Event Sequence Data
[3] Adrian Benton, Margaret Mitchell, and Dirk Hovy. Multi-task learning for mental health using social media text. *arXiv*, 1712.03538, 2017.
[4] Mihajlo Grbovic, Haibin Cheng Real-time Personalization using Embeddings for Search Ranking at Airbnb. KDD 2018