RESEARCH ARTICLE OPEN ACCESS

Review on Disease Prediction Via Graph Neural Networks

Mr. Umesh A. Patil (Assistant Professor), Miss Roshni Rajendra Mahutkar (Student),
Miss Ekta Baburao Mohite (Student), Mr. Devraj Ranjit Mohite (Student), Mr. Vijaykumar Mohan Shete (Student)
(Department of Computer Science And Engineering, D. Y. Patil Technical Campus Talsande, Kolhapur)

ABSTRACT
Electronic Medical Records (EMR) has increasingly available, so disease prediction has recently gained immense research attention, where an accurate classifier needs to be trained tomap the input prediction signals to the estimated diseases for each patient. In this paper, we introduce a model based on Graph Neural Networks (GNNs) for disease prediction, which utilizes external knowledge bases to augment the insufficient EMR data, andlearns highly representative node embeddings for patients, diseases, and symptoms from the medical concept graph and patient record graph.

By aggregating information from directly connected neighbor nodes, the neural graph encoder can effectively generate embeddings that capture knowledge from both data sources and can inductively infer the embeddings for a new patient based on the symptoms reported inpatient EMRs to allow for accurate prediction on both general diseases and rare diseases

Keywords - Disease Prediction, Big Data Health Applications, Data Mining, Graph Embedding

Date of Submission: 28-04-2022 Date of Acceptance: 09-05-2022

I. INTRODUCTION

AS a widely-used data management scheme, electronic medical records (EMRs) are used to store the rich clinical data collected from different patients' visits to hospitals. Recently, with the prosperous advances in information technology and machine learning, the sheer volume of EMRs is becoming more manageable, and analyzing EMRs with machine learning and data mining techniques is becoming an emerging research direction to fulfill the goal of improving health care services. We identify the challenges of predicting both common and rarediseases based onEMRs
and introduce a systematic solution by fusing expert knowledge with machine learning techniques
The contributions of this paper are summarized as follows.

- We identify the challenges of predicting both common and rare diseases based on EMRs and introduce a systematic solution by fusing expert knowledge with machine learning techniques.
- We propose a novel graph embedding-based model for disease prediction. The model inductively learns embeddings from the medical concept graph and patient record graph respectively extracted from the external knowledge base and EMR data, while being able to handle new patients and identify highly relevant symptoms to support accurate disease prediction.
- Extensive experiments on real- world EMR datasets have been conducted, and theresults suggest that our model.

II. PROBLEMSTATEMENT

(Rare)Disease prediction:
Given a medical concept graph C = (Vm ⍰Vs, εMS), a patient record graph P = (Vp ⍰V's, εPS) and the corresponding labels, our goal is to learn a Graph Convolutional Network-based model, which can predict the diseases for each new patient p Ɇ VP. Apart from general disease prediction, by evaluating the prediction performance exclusively on patients who are diagnosed with rare diseases in the real clinical dataset

III. OBJECTIVES:

- To develop the correlated graph for the systematic solution to predicting both common and rare diseases.
- To Propose a novel embedding- based model for disease prediction.
- To find the accuracy of patient disease using GNN.

IV. PROPOSEDWORK:

1. Medical Concept Graph and Patient Record Graph Module:

Medical Concept Graph:

For the particular disease, its associated medical concepts like name in diagnostic symptoms and category belong to it as the medical concepts various in different medical knowledge bases without losing principles we only consider the common concept. For example, symptoms in the paper

We represent these medical concepts as the medical concepts graph which is denotedby,

C= (Vm U Vs, ∑ms)

where,

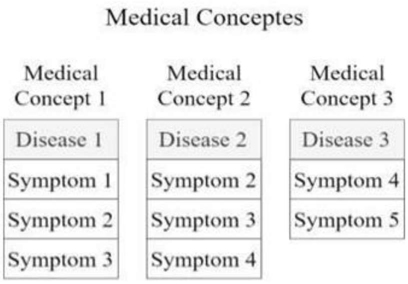
Vm= Nodes of disease Vs= Node set ofsymptoms

∑m= Edge Set

If a symptom s ε Vm is associated with m ε Vm

The construction process of the medical concepts graph is depicted in fig (a) Each disease is the m ε Vm has a |Vm| - dimensional one-hot encoding withthe

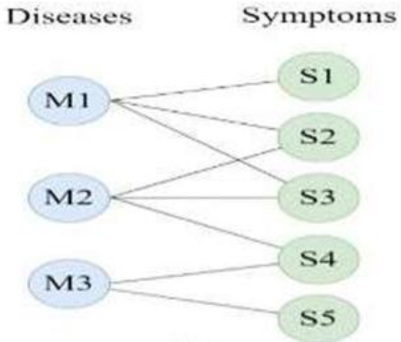
m-th positions it is uniquelyidentified.



Medical Conceptes

Medical Concept 1	Medical Concept 2	Medical Concept 3
Disease 1	Disease 2	Disease 3
Symptom 1	Symptom 2	Symptom 4
Symptom 2	Symptom 3	Symptom 5
Symptom 3	Symptom 4	

(a)



(b)

Patient Record Graph:

We represent the patient record as the patient record graph which is denoted by P = Vp U Vs, ∑ps

where,

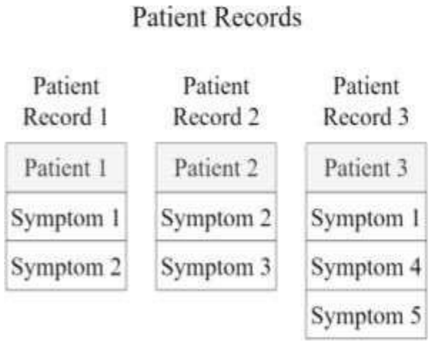
P = Graph structured representation of the EMRs

Vp= The set of nodes of all patient

V's = All the symptoms occurred in EMRs

∑ps= All observed edges between patient node and symptoms node.

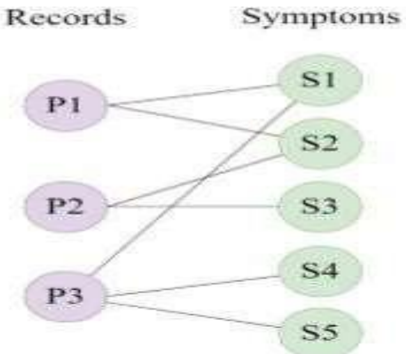
An example is shown in fig(c) and fig(d).,



Patient Records

Patient Record 1	Patient Record 2	Patient Record 3
Patient 1	Patient 2	Patient 3
Symptom 1	Symptom 2	Symptom 1
Symptom 2	Symptom 3	Symptom 4
		Symptom 5

(c)



(d)

Each record P ε Vp is assigned a multi-hot encoding Cp ε {0 ,1} indicating the disease this patient has corresponding indexes are marked as 1.

2. Graph Neural Module using meanAggregator Function:

The figure shows the workflow of our proposed framework for disease prediction. Thismodel takes the medical concept graph C and patient record graph P as its input then embeds every node in both graphs by aggregating the information from their sampledneighbors.

For a given patient, we can form a vector representation by fusing the learned embeddings of symptoms described in EMRs. Then, by measuring the closeness between the embeddings of the patient and any disease, we can eventually estimate the likelihood of diagnosing patient p with disease m.

In this section, we first describe a graph encoder model, which is responsible for producing a low-dimensional embedding z∈⍰⍰for each node in an arbitrary graph.

To begin with, for a graph G = {C, P}, we uniformly represent a disease, symptom, or patient node as v ∈ G to be succinct. Then, at the l-th information propagation layer, the embedding h l v of node v is calculated as:

$$\mathbf{h}^{l}_{\mathcal{N}(v)} = AGGREGATE(\{\mathbf{h}^{l-1}_{v'}, \forall v' \in \mathcal{N}(v)\})$$

$$^{l}_{v} = \sigma(\mathbf{W}^{l} \cdot [\mathbf{h}^{l-1}_{v}; \mathbf{h}^{l}_{\mathcal{N}(v)}])$$

where Wl is the weight matrix to be learned at the l-th layer, h l−1 v is node v's embedding at the previous layer, and we denote the total layer size as L.

We use [·; ·] to represent the concatenation of two vectors, and use N (v) to denote the set of evenly sampled neighbor nodes of v.h 0 v will be initialized as a real-valued dense feature vector, and each digit in h 0 v represents the observed value of a feature dimension. (. h l N(v)) is the synergic representation resulting from the aggregation function. σ is a non-linear activation function (e.g., tanh), and the aggregator can be chosen as mean, max pooling, RNNs.

JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS

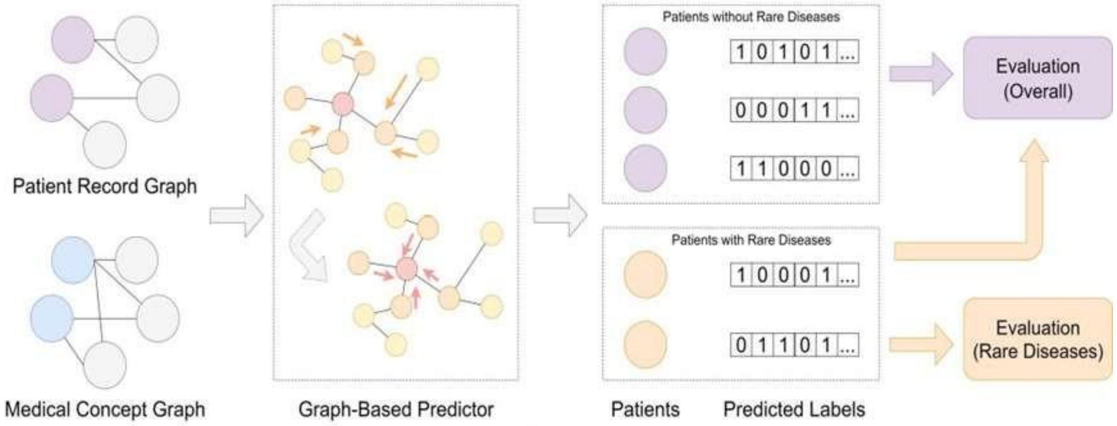


Fig. 2. The workflow of our graph neural network-based model for disease prediction.

The normalization step before reaching the final embedding for all nodes at the last layer

$$\mathbf{h}_{v} = \frac{\mathbf{h}^{L}_{v}}{||\mathbf{h}^{L}_{v}||_2}, \ \forall v \in \mathcal{G}$$

It is worth mentioning that the representations learned for the symptom nodes p in boththe medical concept graph C and patient record graph P share the same embedding space. That is to say, for each symptom p, its embeddings remain the same in both graphs, thus serving as an effective bridge between patient and disease nodes from different graphs. Meanwhile, as different types of nodes, i.e., diseases, patients, and symptoms are learned from three separate embedding spaces, we further align their contexts by projecting all node embeddings onto the same space, followed by a non-linear activation:

zv = σ(Whv), ∀v ∈ G

where W is the learnable projection weight, and zv is the final embedding for node v.

3. Module Training with Graph Encoder Kernels Module:

To understand the effectiveness of different neural architectures in disease prediction, we introduce two variants of the graph encoder. We use two encoder architectures the Graph Attention Network [12] and the Graph Isomorphic Network.

Graph Attention Networks (GATs).

GATs apply attention mechanisms to selectively encode the information from neighbors according to their importance to the target node v. This is achieved by takinga weighted sum of the representations of all v's neighbor nodes:

$$\mathrm{h}^{l}_{v} = \sum_{v0\in \mathrm{N}(v)} \alpha_{v'v}\mathbf{M}\mathbf{h}^{l-1}_{v'}$$

where M is the transformation weight matrix, and αv 0v is the attentive weights indicating the importanceofnodev0∈N(v)whencalculatingh l v . Each av 0v is computed via the following attention network

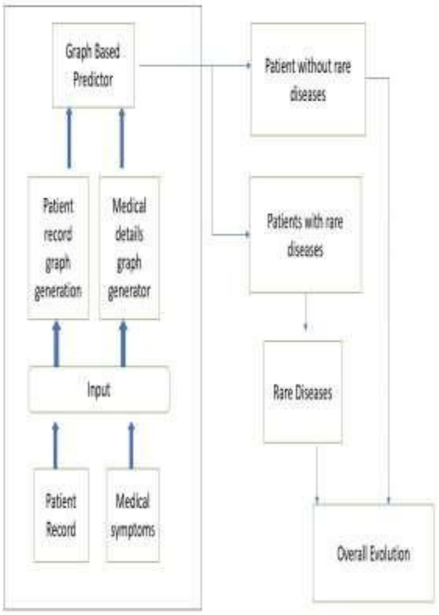
$$\alpha_{v}0_{v} = \frac{\exp(\mathrm{LeakyReLU}(\mathrm{a}^{T}[\mathbf{Nh}^{l-1}_{v}||\mathbf{Nh}^{l-1}_{v'}]))}{\sum k\in \mathrm{N}(v)\ \exp(\mathrm{LeakyReLU}\ (\mathrm{a}^{T}[\mathbf{Nh}^{l-1}_{v}||\mathbf{Nh}^{l-1}_{v'}]))}$$

Graph Isomorphic Networks (GINs):
GINs are claimed effective in representing isomorphic and non-isomorphic graphs with discrete attributes. The computation process of its l-th layer is defined as:

$$\mathrm{h}_v^l = MLP(\sum \mathbf{h}_u^{l-1})$$

where MLP is a multi-layer perceptron. In contrast to other graph encoder kernels which combine the information from both node v itself and its neighbors, a GIN-based graph encoder forms the embedding for v purely based on the embeddings of neighbor nodes.

V. ARCHITECTURE:


Graph Based Predictor
Patient without rare diseases
Patient record graph generation
Medical details graph generator
Patients with rare diseases
Input
Rare Diseases
Patient Record
Medical symptoms
Overall Evolution

VI. IMPLEMENTATION STEPS:

DATASET:
We are using the Proprietary EMR dataset for constructing the patient record graph, whichis our private real-world patient clinical record dataset collected from local hospitals.

MODULES:

- Medical Concept Graph and Patient Record Graph Module
- Graph Neural Module using mean Aggregator Function Module
- Module Training with Graph Encoder Kernels Module

REQUIREMENT SPECIFICATION:
Hardware Requirements:

- HDD:4GB
- OS: Windows Family Software Requirements:
- Python
- Jupyter Notebook
- MYSQL

REFERENCES:

[1]. Q. Suo, F. Ma, Y. Yuan, M. Huai, W.Zhong, Zhang, and J. Gao, "Personalized disease €prediction using a CNN-based similarity learning method," in BIBM. IEEE Computer Society, 2017, pp. 811–816.

[2]. F. Ma, J. Gao, Q. Suo, Q. You, J. Zhou, and A. Zhang, "Risk predictionon electronic healthrecords with prior medical knowledge," in SIGKDD, 2018, pp. 1910–1919.

[3]. H. J. Dawkins, R. Draghia-Akli et al., "Progress in rare diseases research 2010–2016: an irdirc perspective," Clinical and translational science, vol. 11, no. 1, p. 11, 2018.