Modern Management based on Big Data II and Machine Learning and Intelligent Systems III
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA210279

Causal Reasoning Model Based on Medical Knowledge Graph for Disease Diagnosis

Ze XU [a], Huazhen WANG [a, 1], Xiaocong LIU [a], Ting HE [a] and Jin GOU [a]
[a] *College of Computer Science and Technology, Huaqiao University, Xiamen 36102, China*

Abstract. In view of the non-interpretability of disease diagnosis models based on deep learning, a knowledge reasoning model based on medical knowledge graph for intelligent diagnosis is proposed. Given the patient symptom set, the co-occurrence of the patient and the disease is calculated, then the patient suffering from one disease is calculated. Based on the dynamic threshold value, the final disease diagnosis result of the patient is outputted. According to the symptoms of patients and the symptoms in the knowledge graph, the causal reasoning of the disease diagnosis is interpretable. Experiments on 145,712 pediatric electronic medical records in Chinese show that the proposed model can predict diseases with interpretability, and the accuracy reaches-82.12%.

Keywords. Knowledge graph, knowledge reasoning, disease diagnosis, Causal reasoning

1. Introduction

With the development of artificial intelligence technology, disease intelligent diagnosis based on mass medical data has become an important research content in the field of medical information [1]. Medical data, including electronic medical records (EMR), medical dictionaries, medical guidelines and genetic data, mainly exist in the form of text. In clinical practice, the diagnosis of diseases is usually based on patients' symptoms and examination results, combined with doctors' expertise, which is the most important decision support. Therefore, when using artificial intelligence technology to diagnose disease, it is natural to think of building medical professional knowledge into a medical knowledge graph to support medical decision-making [2,3]. The knowledge graph represents the entities and their relationships in a network structure, which can more clearly and intuitively mine the hidden knowledge [4]. In recent years, the research of disease diagnosis based on medical knowledge graph reasoning has been a hot spot in the application of medical artificial intelligence.

According to research, the methods of disease reasoning can be divided into two categories. The first method uses neural network to learn the representations of knowledge graph, which are used for realizing knowledge reasoning [5]. This method can automatically learn the feature of the data without the participation of experts, but because of the lack of interpretability, it is difficult to be widely used in clinical practice.

[1] Corresponding author: Huazhen Wang, College of Computer Science and Technology, Huaqiao University, Xiamen 36102, China; E-mail: wanghuazhen@hqu.edu.cn.

The second one is based on rules and logic, which gives comprehensive rules and ontology to obtain the result of knowledge reasoning, which is essentially logical knowledge reasoning.

Intelligent disease diagnosis is a knowledge-intensive application, and expert knowledge plays a key role in diagnosis decision. Therefore, making full use of the medical knowledge graph and logical reasoning method can obtain the disease reasoning results with high interpretability and accuracy. In this paper, we constructed a knowledge reasoning model based on causal reasoning for disease diagnosis task on real EMR datasets.

2. Related Work

2.1 Knowledge Reasoning Over Knowledge Graph

Knowledge reasoning is mainly focused on analyzing data and finding or inducing knowledge. In the aspect of knowledge reasoning, the most useful method is logical knowledge reasoning and the method based on neural network [6]. In terms of knowledge, there are many forms of knowledge, such as traditional tables or triples of knowledge graph. Knowledge graph, as a semantic network to reveal the relationship between entities, can formally describe the things and their relationships in the real world. discovering the connections between different things. Using knowledge graph as the source of knowledge reasoning can better assist intelligent system to discover knowledge.

With the emergence of knowledge graph, knowledge reasoning over knowledge graph has attracted more and more attention for serving intelligent system [7]. In the field of medicine, there are many medical knowledge graphs, such as Bio2RDFand Liked Life Data [8]. Shi et al. [9] constructed a graph knowledge from medical texts, and carried out semantic reasoning in this a graph knowledge to realize disease diagnosis. But this method ignores the causal relationship between disease and symptoms, which plays an important role in disease diagnosis. Therefore, on the basis of knowledge reasoning, we apply causal reasoning method to the intelligent diagnosis of disease. Specifically, according to the symptoms of patients and the symptoms in the knowledge graph, the disease of patients can be inferred.

2.2 Intelligent Diagnosis of Related Diseases

The emphasis of computer-aided diagnosis is to use information technology to assist disease diagnosis [10]. Early researchers mainly relied on the diagnostic rules made by medical experts for auxiliary diagnosis. With the development of machine learning technology, researchers used Naive Bayesian algorithm and Support Vector Machine algorithm to build disease diagnosis model [11, 12]. In recent years, deep learning methods such as convolutional neural network (CNN) and RNN have been paid more and more attention by related researchers [13, 14]. Dong et al. [15] integrated the information of medical knowledge graph on the basis of deep learning method, so as to improve the accuracy of diseases diagnosis. However, due to lack of interpretability, these methods cannot be widely applied to clinical tasks. In this work, we propose a knowledge reasoning model based on medical knowledge and causal reasoning, trying to provide high interpretability and accuracy.

3. The Proposed Model for Disease Diagnosis

In this paper, we present a disease diagnosis model based on knowledge graph and causal reasoning, that is, according to the symptoms of patients and the symptoms in the knowledge graph, the disease of patients can be deduced, which is interpretable.

The essence of disease is symptoms, and symptoms are a phenomenon of disease. Patients may suffer from multiple diseases at the same time, and different diseases may not occur independently. The symptoms caused by the disease may not all show up. According to the patient's symptom set $suArray$, we can retrieve the symptoms $svArray$ that do not appear on the patient but exist in the medical knowledge graph. According to the patient's symptoms $suArray$, we can retrieve the diseases associated with them in the knowledge graph, as well as the symptoms $suArray$ of these diseases that do not appear in patients but exist in the medical knowledge graph. Then $suArray$ and $svArray$ were used to infer the patient's diseases. The implementation roadmap of the model is shown in Figure 1.

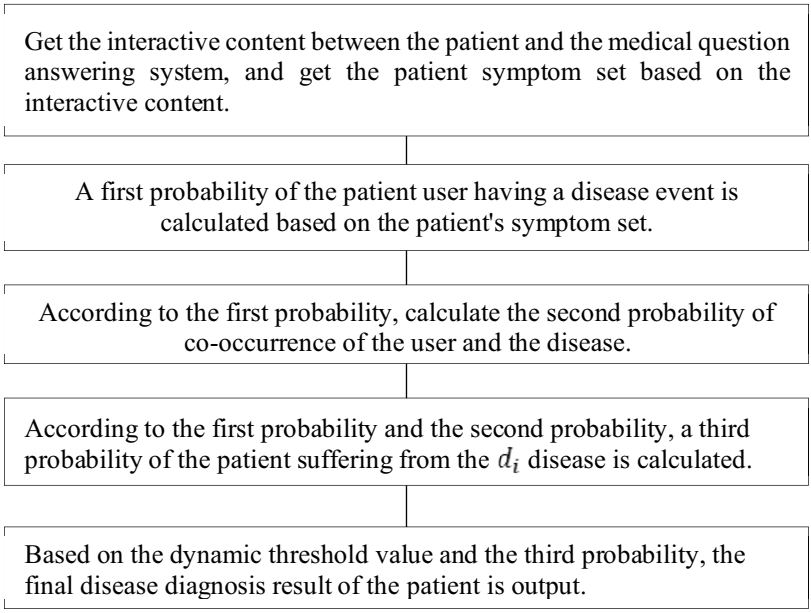

Figure 1. Implementation roadmap of the proposed model.

Let the set of diseases be $\mathrm{D} = \{d_1, d_2, \ldots d_i \cdots, d_n\}$, $1 \ll i \ll n$. Let the set of symptoms be $\mathrm{S} = \{s_1, s_2, \cdots s_j, \cdots, s_m\}$, $1 \ll j \ll m$. The event $\mathcal{D}_\mathrm{i}$ is that the diseases in the collection $\mathrm{D_i}$ occurs simultaneously and none of the disease in set $D - D_i$, where $\mathrm{D_i}$ is a subset of the disease set D, and D_i is any one of 2^D, that is $D_i \in \{D_0, D_1, \cdots, D_{2^n-1}\}$. At first, the first probability $P(\mathcal{S})$ that patient $\mathcal{S}$ has event $\mathcal{D}_i$ is calculated, and its formula is:

$$P(\mathcal{S}) = \sum_{k=0}^{2^n-1} P(\mathcal{D}_i) P(\mathcal{S}|\mathcal{D}_i) \tag{1}$$

When $P(\mathcal{S})$ is expanded according to the full probability formula, since $d_1, d_2, d_3, \cdots, d_n$ is not mutually exclusive events, it cannot constitute $P(\mathcal{S})$ complete event graph. Set $\mathcal{D}' \subset D$ makes the diseases in $\mathcal{D}'$ constitute $P(\mathcal{S})$ event graph. $P(\mathcal{D}')$ is the probability of the complete event graph $\mathcal{D}'$, and $P(\mathcal{S}|\mathcal{D}')$ is the probability of the complete event graph $\mathcal{D}'$ under patient $\mathcal{S}$.Then, the second probability $P(d_i\mathcal{S})$ of patient S and disease d_i is calculated, and the formula is:

$$P(d_i\mathcal{S}) = P\left(\left(\bigcup_{d_i \in D'} \mathcal{D}'\right)\mathcal{S}\right) = P\left(\bigcup_{d_i \in D'} \mathcal{D}'\mathcal{S}\right) = \sum_{d_i \in D'} P(D'\mathcal{S})$$

$$= \sum_{d_i \in D'} P(\mathcal{D}')P(\mathcal{S}|\mathcal{D}') \tag{2}$$

Obtaining the probability $P(\mathcal{S})$ of the existence of the patient $\mathcal{S}$ and the probability $P(d_i\mathcal{S})$ of the co-occurrence of the patient $\mathcal{S}$ and the d_i disease, and calculating the third probability $P(d_i|\mathcal{S})$ of the patient $\mathcal{S}$ suffering from the d_i disease, wherein the formula is as follows:

$$P(d_i|\mathcal{S}) = \frac{P(d_i\mathcal{S})}{P(\mathcal{S})} = \frac{\sum_{d_i \in D'} P(\mathcal{D}')P(\mathcal{S}|\mathcal{D}')}{\sum_{k=1}^{2^n-1} P(\mathcal{D}_i)P(\mathcal{S}|\mathcal{D}_i)} \tag{3}$$

In the next part, a final disease diagnosis result of the patient $\mathcal{S}$ is output based on the dynamic threshold and the third probability. The specific steps are as follows:

First of all, arrange the diseases in descending order according to the probability value obtained previously, to obtain the descending disease probability list O_r with the probability greater than 0, that is, $Q_r = [disease\ 1{:}\ probability\ 1, disease\ 2{:}\ probability\ 2, \ldots]$;

Then, the output threshold λ is set, and the probability of the descending disease probability list is compared with the set threshold λ. The disease whose probability is greater than or equal to the set threshold is constructed into the selected disease list O_{re} and output directly. The disease whose probability is less than the set threshold is constructed into the unselected disease list O_{rne} for the next step of screening.

According to $\mathcal{D}'$, $\mathrm{P}(\mathcal{D}')$ and the probability $\mathrm{P}(s_j|\mathcal{D}')$ of symptom s_i occurring under the condition of complete event graph $\mathcal{D}'$, Calculate the fourth probability $\mathrm{P}(s_j)$ of the occurrence of the unmanifested symptom s_j, and its formula is:

$$P(s_j) = \sum_{D' \subset D} P(\mathcal{D}')P(s_j|\mathcal{D}') \tag{4}$$

Selecting the non-appearing symptom corresponding to the maximum probability as the most likely appearing symptom to inquire about the patient; If the emerging symptom occurs, the emerging symptom is added to the emerging symptom set of the patient $\mathcal{S}$. The descending disease probability list is recalculated. And then, according to the probability of the occurrence of the disease d_i and the increment $\triangle P(d_i)$ of the

probability of the occurrence of the disease d_i before and after interaction, calculating the increment $\triangle I$ of the uncertainty, wherein the formula is as follows:

$$\triangle I = \sum_{d_i \in D \wedge P(d_i)>0} \triangle I(d_i) = - \sum_{d_i \in D \wedge P(d_i)>0} \log_2 \frac{P(d_i) + \triangle P(d_i)}{P(d_i)} \tag{5}$$

Then, calculate a dynamically adjusted threshold value $\lambda' = \lambda - \triangle \lambda = \lambda - \triangle I$, selected diseases in that unselected disease list Q_{rne} with the probability larger than the current threshold value, and construct an interactive disease list Q_{rei};

Finally, the selected disease list Q_{re} and the interactive disease list Q_{rei} are summarized to form a final disease diagnosis result list Q_{ree} as the final disease diagnosis result.

4. Experimental Setup and Results

4.1 Construction of Medical Knowledge Graph

In this paper, a bottom-up method was adopted to construct the medical knowledge graph. Firstly, we extracted the triplet of "disease" and "symptom" from the several medical websites: http://tag.120ask.com/, http://jbk.39.net/, http://jbk.familydoctor.com.cn/. Secondly, after entity alignment and knowledge fusion, these triples were stored in neo4j database. The final knowledge graph includes 7958 diseases, 5747 symptoms, and 65530 connections between them. An example diagram of the medical knowledge graph is shown in Figure 2.

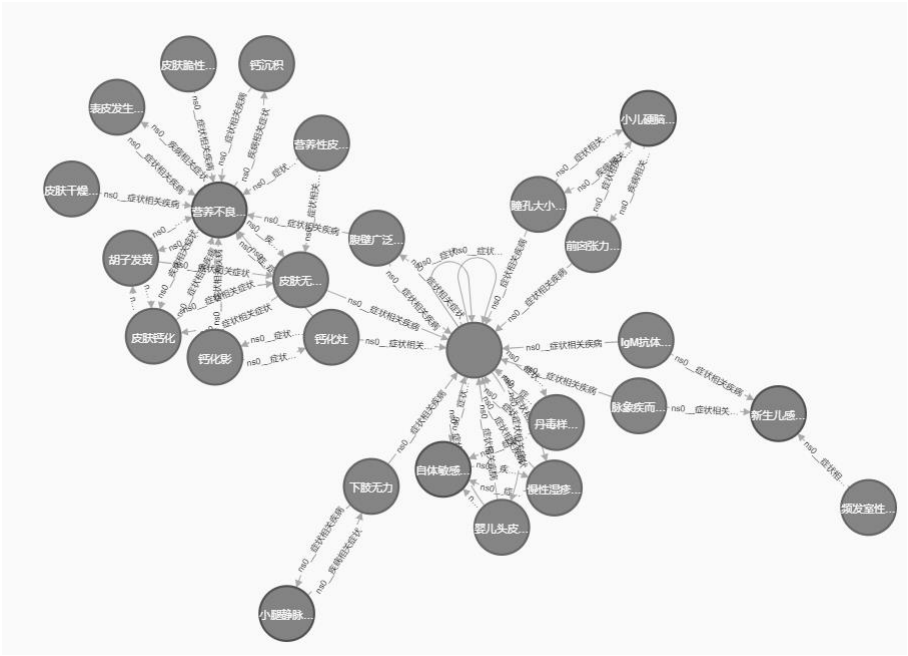

Figure 2. An example diagram of our medical knowledge graph.

4.2 Experimental Results

The dataset used in this paper come from a tertiary hospital, including 145,712 pediatric outpatient medical records, each of which has only one preliminary diagnosis. The purpose of this paper is to input the patient's symptoms and predict the patient's disease. The disease with the highest prediction probability is taken as the result. The evaluation index of the model is calculated as follows:

$$\text{Accuracy} = \frac{I_r}{I} \tag{6}$$

where I_r is the number of samples with correct diagnosis predicted by model, and I is the number of all samples.

We set the model proposed by Li et al. [16] as baseline model, which used deep convolutional neural network for pediatric disease prediction. Table 1 shows the experimental results.

Table 1. Comparison of experimental results between two models

Model	Accuracy
Our model	82.12%
[16]	71.07%

It can be seen from Table 1 that the accuracy of our model reaches 82.12%, which is 11.05% higher than that of CNN model, indicating that the causal reasoning model based on knowledge graph has outstanding performance in disease diagnosis.

5. Conclusion

Disease diagnosis is a prominent clinical application in clinical decision support. In this paper, we propose a knowledge reasoning model based on medical knowledge graph and causal reasoning for intelligent diagnosis. According to the patient's symptoms, we can retrieve the diseases associated with them in the knowledge graph, as well as the symptoms of these diseases that do not appear in patients but exist in the medical knowledge graph, thus inferring the patient's diseases. Then, our model can deduce the patient's disease set based on causal reasoning. Experiments on 145,712 pediatric electronic medical records in Chinese show that the proposed model can predict diseases with interpretability and high accuracy. It can be also applied to medical question and answer systems, medical auxiliary diagnosis and intelligent personal health assistants. Our future work will establish a more complete medical knowledge graph and apply our model to more diseases.

Acknowledgments

Research works in this paper are supported by the National Key Technology R&D Program of China (No.2018YFB1402500), the Social Science Planning Foundation of Fujian Province (FJ2020B0033), and Huaqiao University's Academic Project Supported by the Fundamental Research Funds for the Central Universities (TZYB-202005).

References

[1] Sun XX. The Construction of academic knowledge mapping based on latent semantic analysis. Central China Normal University, 2013.
[2] Zhao J, Liu K, He SZ, et al. knowledge Graph. Chin. J. Info., 2020:34 (09):111.
[3] Giger ML, Chan HP, Boone J. Anniversary paper: history and status of CAD and quantitative image analysis: the role of medical physics and AAPM. Medi Phys, 2008: 35(12):5799-5820.
[4] Dai Y, Wang S, Xiong NN, et al. A Survey on Knowledge Graph Embedding: Approaches, Applications and Benchmarks. Electronics, 2020, 9(5):750.
[5] Wang D, Amriljaharadak A, Xiao Y. Dynamic knowledge inference based on bayesian network learning. Mathematical Problems in Engineering, 2020, 2020(4):1-9.
[6] Guan SP, Jin XL, Jia YT, et al. Research progress of knowledge reasoning for knowledge graph. Acta Software Sinica. 2018, 29(10):74-102.
[7] Hu GY. Research on the construction of knowledge graph of professional knowledge and skill system. Industry and Info. Edu. 2020(12): 123-127.
[8] Jia LR, Liu J, Yu T, et al. Construction of traditional Chinese medicine knowledge graph. J. Med. Info. 2015, 36(08):51-53+59.
[9] Shi L, Li S, Yang X, et al. Semantic health knowledge graph: semantic integration of heterogeneous medical knowledge and services. Biomed Res Int. 2017(2017):2858423.
[10] Shortliffe EH, Axline SG, Buchanan BG, et al. An Artificial Intelligence program to advise physicians regarding antimicrobial therapy. Compute Biomed Res, 1973, 6(6):544-560.
[11] Baati K, Hamdani T M, Alimi AM. Diagnosis of lymphatic diseases using a naive bayes style possibilistic classifier. Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics. IEEE, 2013.
[12] Kunwar V, Chandel K, Sabitha AS, et al. Chronic kidney disease analysis using data mining classification techniques. Cloud Syst. Big Data Engineering. IEEE, 2016.
[13] Tsehay YK, Lay NS, Roth HR, et al. Depth learning architecture based on convolutional neural network for detecting prostate cancer on multi-parametric magnetic resonance images. J. Medical Imaging 2017: Computer-Assisted Diagnosis, Orlando, February 11-16, 2017. San Francisco: SPIE, 2017: 1013405.
[14] Esteva A, Kuprel B, Novoa R A, et al. Corrigendum: Dermatologist-level classification of skin cancer with deep neural networks. Nature, 2017, 546(7660):686-686.
[15] Dong LL, Cheng J, Zhang X. Research on disease diagnosis method integrating knowledge graph and deep learning. Computer Sci. and Explor. 2020, 14(05):815-824.
[16] Li XZ, Wang HZ, Xiong YJ, et al. Application of convolutional neural network in pediatric disease prediction. China Digital Med. 2018,13(10): 11-13.