Knowledge-Based Systems 000 (2017) 1–11

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

# Learning and inference in knowledge-based probabilistic model for medical diagnosis

Jingchi Jiang[a], Xueli Li[b], Chao Zhao[a], Yi Guan[a,*], Qiubin Yu[c]

[a] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
[b] EBAONET Healthcare Information Technology (Beijing) CO.LTD, Beijing 100028, China
[c] Medical Record Room, Second Affiliated Hospital of Harbin Medical University, Harbin 150086, China

ARTICLE INFO

*Article history:*
Received 19 September 2016
Revised 21 September 2017
Accepted 24 September 2017
Available online xxx

*Keywords:*
Probabilistic model
First-order knowledge
Markov network
Gradient descent
Markov logic network

ABSTRACT

Based on a weighted knowledge graph to represent first-order knowledge and combining it with a probabilistic model, we propose a methodology for creating a medical knowledge network (MKN) in medical diagnosis. When a set of evidence is activated for a specific patient, we can generate a ground medical knowledge network that is composed of evidence nodes and potential disease nodes. By incorporating a Boltzmann machine into the potential function of a Markov network, we investigated the joint probability distribution of the MKN. To consider numerical evidence, a multivariate inference model is presented that uses conditional probability. In addition, the weights for the knowledge graph are efficiently learned from manually annotated Chinese Electronic Medical Records (CEMRs) and Blood Examination Records (BERs). In our experiments, we found numerically that an improved expression of evidence variables is necessary for medical diagnosis. Our experimental results comparing a Markov logic network and six kinds of classic machine learning algorithms on the actual CEMR database and BER database indicate that our method holds promise and that MKN can facilitate studies of intelligent diagnosis.



## 1. Introduction

The World Health Organization (WHO) reports that 422 million adults have diabetes, and 1.5 million deaths are directly attributed to diabetes each year [1]. Additionally, the number of deaths caused by cardiovascular diseases (CVDs) and cancer annually is estimated to be 17.5 million and 8.2 million, respectively [2]. The WHO report on cancer shows that the number of new cases of cancer will increase by 70% over the next two decades. In the face of this situation, researchers have begun to pay more attention to health care. According to existing studies, more than 30% of cancer deaths could be prevented by early diagnosis and appropriate treatment [3]. Because an accurate diagnosis contributes to a proper choice of treatment and subsequent cure, medical diagnosis plays a significant role in improving health care. Consequently, a means to provide an effective intelligent diagnostic method to assist clinicians by reducing the costs and improving the accuracy of diagnosis has been a critical goal in the efforts to enhance the patient medical service environment.

Classification is one of the most widely researched topics in medical diagnosis. The general model classifies a set of symptom data into one of several predefined categories of disease for cases of medical diagnosis. A decision tree [4,5] is a classic algorithm in the medical classification domain, one that uses the information entropy method; however, it is sensitive to inconsistencies in the data. The support vector machine [6–8] has a solid theoretical basis for the classification task; because of its efficient selection of features, it has higher predictive accuracy than decision trees. Bayesian networks [9,10], which are based on Bayesian theory [11,12], describe the dependence relationship between the symptom variables and the disease variables; these can be used in medical diagnosis. Other diagnostic models include neural networks (NN) [13–15], fuzzy logic (FL) [16,17], and genetic algorithms (GAs) [18–20]. Each of these is designed with a distinct methodology for addressing diagnosis problems.

The existing studies have mainly focused on exploring effective methods for improving the accuracy of disease classification. However, these methods often ignore the importance of the application of domain knowledge. Although a hybrid Markov logic network (HMLN) [21], which is a generalization of a Markov logic network (MLN) [22], aims to integrate Boolean and numerical variables into a probabilistic logic modeling framework, inference is typically done by estimation methods that are based on variable

* Correspondence address.
*E-mail addresses:* jiangjingchi0118@163.com (J. Jiang), xueli.li@ebaonet.cn (X. Li), guanyi@hit.edu.cn (Y. Guan), yuqiubin6695@163.com (Q. Yu).

https://doi.org/10.1016/j.knosys.2017.09.030
0950-7051/

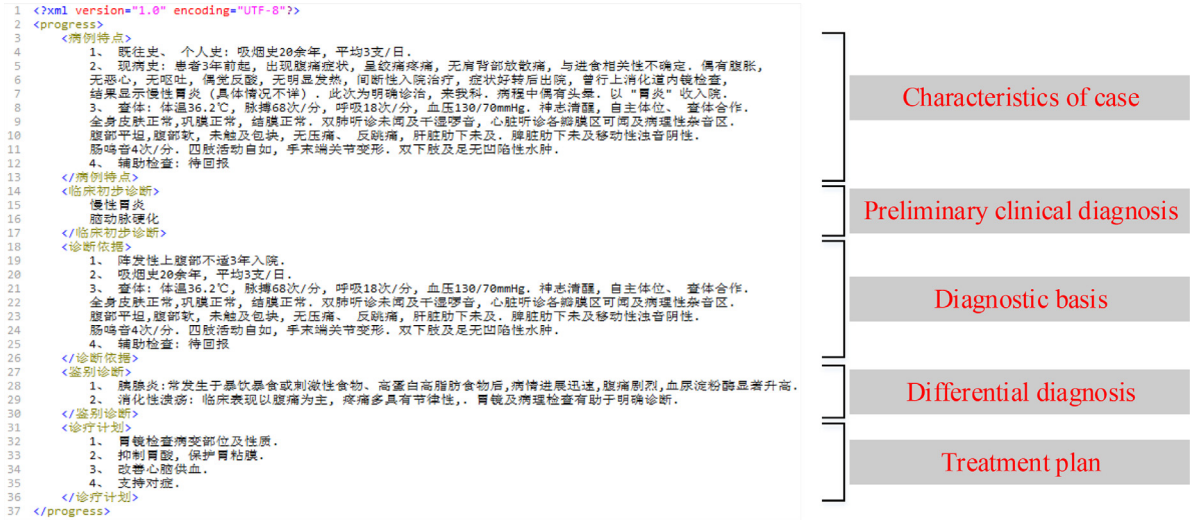


**Fig. 1.** Sample of progress note from the Second Affiliated Hospital of Harbin Medical University.

approximations or sampling strategies [23,24]. In addition, the inference efficiency of MLN and HMLN decreases steadily as the number of constraints increases. This is a significant problem because inference can become intractable for certain types of domain knowledge [25,26], especially in the health care domain, which contains a large amount of medical knowledge. In this paper, we focus on combining medical knowledge with a novel probabilistic model to assist clinicians in making intelligent decisions and show how this novel probabilistic model can be applied in medical diagnosis. We conducted our investigation as follows:

(1) Based on Chinese Electronic Medical Records (CEMRs) and Blood Examination Records (BERs), we developed techniques for the recognition of named entities and entity relationships. According to the relational structure between named entities, we built a symptom–disease knowledge base and an examination–disease knowledge base, each consisting of a set of rules in first-order logic.
(2) We mapped each first-order knowledge base into a knowledge graph. Each graph is composed of first-order predications (nodes) and diagnostic relationships among predications (edges). Furthermore, the graph can also be an intuitive reflection of the inferential structure of the knowledge.
(3) We developed a novel probabilistic model for medical diagnosis that is based on Markov network theory. To adapt it to the requirements of the multivariate feature, we incorporated a Boltzmann machine into the potential function of the Markov network. It can simultaneously model both binary and numerical variables. The mathematical derivation of learning and inference is rigorously deduced.
(4) By a numerical comparison with other diagnostic models for CEMRs and BERs, we found that our probabilistic model is more effective for diagnosing several diseases according to the measure of precision for the first 10 results (P@10) and that of recall for the first 10 results (R@10).

The rest of this paper is organized as follows. In Section 2, we introduce Chinese Electronic Medical Records, Blood Examination Records, and the knowledge graph. In Section 3, we review the fundamentals of Markov networks and Markov logic networks. In Section 4, the knowledge-based probabilistic model based on Markov networks is proposed; then, we demonstrate the mathematical derivation of learning and inference. In Section 5, we further evaluate the effectiveness and accuracy of our probabilistic model for medical diagnosis. Finally, we conclude this paper and discuss directions for future work in Section 6.

## 2. Knowledge extraction and knowledge representation

### 2.1. Chinese electronic medical records and blood examination records

Electronic medical records (EMRs) [27] are a systematized collection of patient health information in a digital format. As the crucial carrier of recorded medical activity, EMRs contain significant medical knowledge [28,29]. Therefore, for this study, we adopted Chinese Electronic Medical Records (CEMRs) in free-form text as the primary source of medical knowledge. These CEMRs, which have had protected health information (PHI) [30] removed, come from the Second Affiliated Hospital of Harbin Medical University, and we obtained the usage rights for research. These CEMRs include five main kinds of free-form text: discharge summaries, progress notes, patient complaints, patient disease histories, and communication logs. Considering the abundance of medical knowledge and the difficulty of Chinese text processing, we chose the discharge summaries and the progress notes as the sources for knowledge extraction. The structures of the progress note and discharge summary are shown in Figs. 1 and 2, respectively.

In contrast to the structure of CEMRs, BERs describe a series of blood test results, such as the levels of hemoglobin, alanine aminotransferase, and hepatitis B virus surface antigen. Most of the blood examination items use numerical evaluation criteria to make up for the shortage of discrete symptom variables in CEMRs. Furthermore, symptom-based preliminary diagnosis and examination-based definitive diagnosis are two essential components of the medical process. Therefore, BERs are important in disease diagnosis because they enable clinicians to make an accurate diagnosis based on the deviation of examination results from their respective normal ranges. In this study, BERs came from XingYi People's Hospital, and we obtained the usage rights for research. The structure of a BER, with one line per examination item, is shown in Fig. 3.

### 2.2. Corpus

The recognition of named entities [31] and entity relationships [32] is an important aspect in the extraction of medical knowledge from CEMRs. Referencing the medical concept annotation guideline and the assertion annotation guideline given by Informatics for Integrating Biology and the Bedside (i2b2) [33], we have drawn

Please cite this article as: J. Jiang et al., Learning and inference in knowledge-based probabilistic model for medical diagnosis, Knowledge-Based Systems (2017), https://doi.org/10.1016/j.knosys.2017.09.030