Artificial Intelligence in Medicine 87 (2018) 49–59

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

EMR-based medical knowledge representation and inference via Markov random fields and distributed representation learning

Chao Zhao [a], Jingchi Jiang [a], Yi Guan [a,*], Xitong Guo [b], Bin He [a]

[a] School of Computer Science and Technology, Harbin, Heilongjiang 150001, China
[b] School of Management, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China

ARTICLE INFO

Article history:
Received 22 May 2017
Received in revised form 28 February 2018
Accepted 29 March 2018

Keywords:
Electronic medical record
Clinical decision support
Medical knowledge network
Markov random field
Distributed representation

ABSTRACT

Objective: Electronic medical records (EMRs) contain medical knowledge that can be used for clinical decision support (CDS). Our objective is to develop a general system that can extract and represent knowledge contained in EMRs to support three CDS tasks—test recommendation, initial diagnosis, and treatment plan recommendation—given the condition of a patient.
Methods: We extracted four kinds of medical entities from records and constructed an EMR-based medical knowledge network (EMKN), in which nodes are entities and edges reflect their co-occurrence in a record. Three bipartite subgraphs (bigraphs) were extracted from the EMKN, one to support each task. One part of the bigraph was the given condition (e.g., symptoms), and the other was the condition to be inferred (e.g., diseases). Each bigraph was regarded as a Markov random field (MRF) to support the inference. We proposed three graph-based energy functions and three likelihood-based energy functions. Two of these functions are based on knowledge representation learning and can provide distributed representations of medical entities. Two EMR datasets and three metrics were utilized to evaluate the performance.
Results: As a whole, the evaluation results indicate that the proposed system outperformed the baseline methods. The distributed representation of medical entities does reflect similarity relationships with respect to knowledge level.
Conclusion: Combining EMKN and MRF is an effective approach for general medical knowledge representation and inference. Different tasks, however, require individually designed energy functions.

1. Introduction

The purpose of a clinical decision support system (CDSS) is to provide clinicians or patients with computer-generated clinical knowledge and patient-related information that can be intelligently filtered or presented at appropriate times, with the goal of enhancing patient care [1]. A core component of a CDSS is the knowledge base, which was originally established and updated manually by clinical experts but today is increasingly likely to be generated and managed automatically. This process often includes natural language processing (NLP) techniques for mining clinical knowledge from medical free text [2], such as medical literature and electronic medical records (EMRs). Focusing on the former, the 2015 Text REtrieval Conference (TREC) Clinical Decision Support (CDS) track expected to develop a retrieval-based system to solve the following three problems by returning the most relevant biomedical articles [3]:

- Determining a patient's most likely diagnosis given a list of symptoms;
- Determining the most effective treatment plan for a patient with a known condition;
- Determining whether a particular test is indicated in a given situation.

We suggest, however, that it is possible to reduce the granularity of information from articles to medical entities. The CDSS can directly provide the proper test items, list of potential diseases, and sequential treatment plan orders appropriate for the patient condition instead of simply the relevant literature, which the clinician must peruse to find the necessary information. The EMR is a credible source of medical knowledge for this purpose—it contains all of a patient's healthcare data and medical history in an electronic format. These data include abundant medical entities, such as the

* Corresponding author.
E-mail addresses: zhaochaocs@gmail.com (C. Zhao), jiangjingchi0118@163.com (J. Jiang), guanyi@hit.edu.cn (Y. Guan), xitongguo@hit.edu.cn (X. Guo), hebin_hit@hotmail.com (B. He).

https://doi.org/10.1016/j.artmed.2018.03.005
0933-3657/

current clinical diagnosis, medical history, results of lab tests, and treatment plans [4,5]. These entities, and the relationships between them, are the primary carriers of medical knowledge in EMRs and can be automatically extracted using information extraction techniques [6–8].

After the information is extracted, the two subsequent key problems are (1) representing medical knowledge using these entities and their interrelationships and (2) making medical inferences according to this representation. Several machine learning solutions for these two problems have been proposed [9], such as statistical classifiers, association rules, and Bayesian networks. Advances in representation learning and deep learning on NLP have also provided a new approach to CDS. However, because of limitations inherent in the model itself or to the computational complexity, most methods focus only on a specific disease, and universal support systems for general practice remain elusive. We believe that it is important to construct such a general system, for two main reasons. First, such a system could respond to the needs of ordinary people suffering from any problematic symptoms. They may wish to have some idea about their health condition before consulting a doctor. Second, a general system could yield initial results that may be better than those from specialized CDS systems in supporting real-world applications, especially for the general practice. Specialized systems have higher precision requirements and therefore need patient information that is much more detailed than his or her complaint symptoms to guarantee precise results.

In this paper, we make a preliminary attempt to represent medical knowledge from EMRs to resolve the three problems outlined above at a medical entity level. We first represent the medical knowledge using an EMR-based medical knowledge network (EMKN) and then treat it as a Markov random field (MRF) for inference tasks. In the EMKN, the nodes are medical entities, and the edges are entity co-occurrence relationships. The MRF describes the probability distribution among these entities and makes probabilistic inferences according to predefined energy functions.

Our main contribution is threefold:

- We propose a universal EMR-based clinical decision support method using EMKN and MRF. This method takes only the corresponding medical entities as inputs and is not restricted to particular diseases.
- We derive a learning algorithm based on derivable energy functions and integrate the knowledge representation learning approaches into the MRF to obtain a distributed representation of medical entities.
- We apply the inference architecture to three CDS tasks: test suggestion, initial diagnosis, and treatment plan suggestion. This allows us to experimentally demonstrate the efficiency of the proposed method with actual clinical records.

The remainder of this paper is structured as follows. In Section 2, we give a brief review of existing CDS systems as well as work related to representation learning. In Section 3, we describe the details of the construction of the EMKN as well as the inference and learning algorithms based on the MRF. Section 4 introduces two distributed representation methods for representing medical entities as an MRF. We evaluated these methods using two actual EMR datasets in terms of three metrics, as described in Section 5. The results are discussed at length in Section 6. A brief conclusion and a discussion of future research directions are presented in Section 7.

2. Related work

The three problems referred to in Section 1 can be further generalized as a single problem: to provide the best possible clinical recommendations (medical investigations, possible diagnosis, and treatment plans) for a given patient's condition. This section introduces previous work on medical knowledge representation and decision making for this problem.

2.1. Traditional CDS systems

Many of the existing CDS systems focus on one (or one kind of) disease and adopt classification strategies to solve this problem. With these systems, some typical patient features (e.g., signs, symptoms, test results) are extracted with the help of domain experts' knowledge and then transformed and selected by the algorithms. After the feature engineering, the disease condition can be determined by general classifiers such as logistic regression [10,11], neural networks [12,13], or the naïve Bayes [14]. For example, Lasserre et al. [10] constructed a set of classifiers to predict the estimated glomerular filtration rate (eGFR) in kidney transplant patients with the help of 56 selected features from the donor and the recipient.

Other researchers have attempted to develop models that do not require the manual input of prior knowledge to learn the relationships among clinical events directly from the data. Association rules mining is a typical approach for identifying associations of clinical event pairs [15–17]. Bayesian networks (or more generally, "probabilistic graphical models (PGMs)") can also be used to represent the relationships between medical events [18–20]. For example, Klann et al. [20] used a Bayesian network to implement an adaptive recommendation system to display a menu of recommended next treatments based on the sequence of previous treatments. In contrast to association rules mining, Bayesian networks can account for transitive associations and co-varying relationships among variables.

For attempting to diagnose more than one disease with the above approaches, the size of the feature set and the number of random variables would be excessive; thus, neither binary classifiers nor Bayesian networks are a good choice. The former would suffer from the curse of dimensionality and class imbalance, and the latter is limited by the computational complexity of inference and learning [21]. In the worst case, both the inference and structure learning of Bayesian networks are NP-hard [22,23]. To improve the scalability of Bayesian networks on high-dimensionality data, researchers must customize the network by introducing some prior restrictions on the graph structure or the distribution of random variables [24,25], e.g., the hierarchical latent class models [26] or interval graphs [27] in genome-wide association studies. Another commonly used PGM, MRFs, has similar problems but is less often utilized for CDS. In this work, we constrain the MRF structure as a bipartite graph to reduce the computational complexity.

2.2. Representation learning for healthcare

Recent developments in representation learning and deep learning have created new opportunities for medical knowledge representation. The aim of representation learning is to learn a good representation of the data, which can make it easier to extract useful information when building classifiers [28]. One widely used representation technique is deep learning [29], which has been particularly successful in a variety of artificial intelligence (AI) fields [30,31], including NLP [32–34]; for example, researchers attempted to map words to a low-dimensional, dense vector space. The different elements in the vector express the word's features from various perspectives. This method, called *distributed representation*,