Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

Diagnosis labeling with disease-specific characteristics mining

Jun Guo[a,*], Xuan Yuan[a], Xia Zheng[b], Pengfei Xu[a], Yun Xiao[a], Baoying Liu[a,*]

[a] *School of Information Science and Technology, Northwest University, Xian 710127, PR China*
[b] *Department of Culture Heritage and Museology, Zhejiang University, Hangzhou 310028, PR China*

ARTICLE INFO

Keywords:
ICU medical data minings
Diagnosis labeling
Disease-specific features

ABSTRACT

Data analysis and management of huge volumes of medical data have attracted enormous attention, since discovering knowledge from the data can benefit both caregivers and patients. In this paper, we focus on learning disease labels from medical data of patients in Intensive Care Units (ICU). Specifically, we extract features from two main sources, medical charts and notes. We apply the Bag-of-Words (BoW) model to encode the features. Different from most of the existing multi-label learning algorithms that take correlations among diseases into consideration, our model learns disease specific features to benefit the discrimination of different diseases. To achieve this, we first construct features specific to each disease by conducting clustering analysis on its positive and negative instances, and then perform training and testing by querying the clustering results. Extensive experiments have been conducted on a real-world Intensive Care Units (ICU) database. Evaluation results have shown that our proposed method has better performance against all other compared multi-label learning methods.

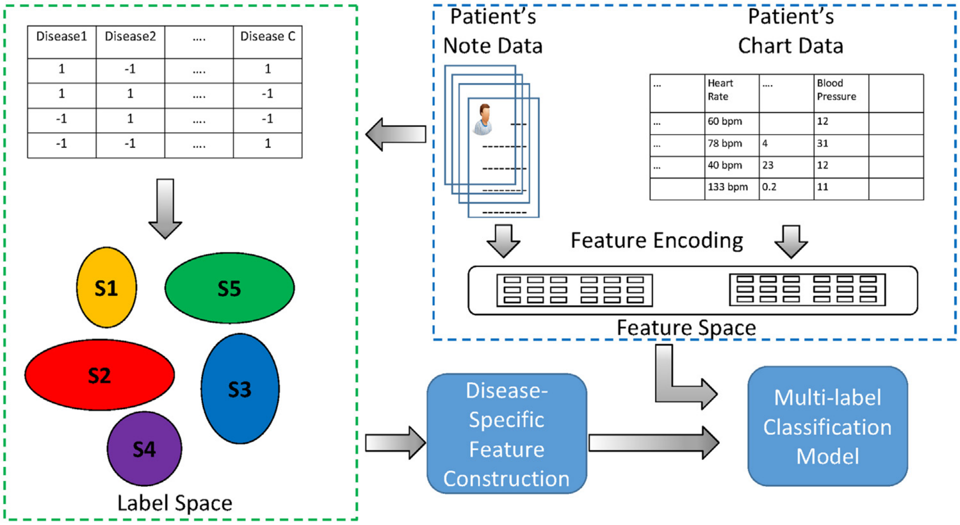
1. Introduction

With rapid developments of medical information system, the amounts of medical data, e.g. Electrical Health Records (EHRs), are booming. These medical data carry with valuable knowledge that is beneficial to medical research. Over the past few decades, researchers in the field of data mining have been paying much attention to uncovering useful patterns from the data. Given such huge amounts of data, assigning labels to the records will facilitate further processing operations, such as indexing, retrieval, etc. For example, when a doctor is making a treatment plan for a patient who suffers from diseases which the doctor has limited experiences. An intuitive solution is to search similar cases using the disease names or codes in a medical database. The returned information that includes all the records of many similar cases supports the caregiver to make better decisions about the treatment. One of the cornerstone components is the disease code system that classifies diseases into different groups. The most widely used disease code system is International Statistical Classification of Diseases and Related Health Problems (usually abbreviated as ICD). ICD is proposed and periodically revised by the World Health Organization (WHO), and its latest version is ICD-10. In many regions in the world, ICD is further modified to fit the local medical systems, e.g. ICD-10-AM for Australia and ICD-9-CM for the United States. The application of ICD code massively facilitates the management of patient records. Therefore, accurate and complete disease labeling plays a significantly important role as the ground truth information in the medical applications.

Conventionally, disease codes are manually labeled by medical experts, i.e. doctors, in the hospital. Before assigning the labels, the experts will review the historical records to collect more prior knowledge of the individual health conditions. All the previous records since from the first admission will help the experts make a right diagnosis (choose one or more proper disease codes). Assigning diagnosis codes to a patient can occur during or after the actual health care. Accurate and complete disease labeling relies on two factors: First, useful temporal information can be identified by the experts using their professional experiences. For example, disease development information is quite helpful when diagnosing a potential disease that is developed from the previous health conditions. Second, disease correlation can be recognized by the experts using their knowledge background. Those diseases that have been already identified can improve the performance when diagnosing a potentially correlated disease. For instance, *Hypertensive disease* (ICD9 401-405) are highly correlated with *Other forms of heart disease* (ICD9 420-429). When considering the possibility of *Other forms of heart disease*, the occurrence of *Hypertensive disease* can help to make a right decision. Unfortunately, human experts hardly keep these two factors all the time in mind when facing the big medical data. To free medical professionals from time-consuming and tedious

[*] Corresponding authors.
E-mail addresses: guojun@nwu.edu.cn (J. Guo), paola.liu@nwu.edu.cn (B. Liu).

https://doi.org/10.1016/j.artmed.2018.06.006
Received 17 August 2017; Received in revised form 15 June 2018; Accepted 26 June 2018
0933-3657/



Disease1	Disease2	....	Disease C
1	-1	....	1
1	1	....	-1
-1	1	....	-1
-1	-1	....	1

...	Heart Rate	....	Blood Pressure	
...	60 bpm		12	
...	78 bpm	4	31	
...	40 bpm	23	12	
	133 bpm	0.2	11	

Fig. 1. Workflow demonstration of the proposed framework. The Blue dashed rectangle on the right includes all patient data in two different types, i.e. chart and note data, from which two types of features are extracted, e.g. Latent Dirichlet Allocation (LDA) for the notes. The Bag-of-Words model is used to encode two types of features into a unified representation for each patient. The green dashed rectangle on the left contains the patient label vectors in the label space. For each disease, we construct a disease-specific feature to benefit the discrimination of different diseases. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

medical record reviews, an effective and automatic labeling system is in great demand. From the perspective of machine learning, assigning ICD codes to each patient data is equivalent to learning multiple labels. The aim is to train a classifier that can output proper labels given a set of medical features. To achieve this, a simple solution is to train a set of classifiers, each for a specific disease. However, this method separately considers each disease patterns, failing to consider the disease-specific features. Therefore, disease-specific features should be exploited to benefit the discrimination of different diseases (Fig. 1).

In this paper, we focus on assigning disease codes to large-scale ICU patient data, which is equivalent to a multi-label classification problem. In our work, we extract two types of features from medical chart data and medical note data, respectively. The medical charts can be regarded as structured data while the medical notes are unstructured data. Chart data are manually collected from monitoring devices by medical staff in ICU according to the patient's health conditions. Compared with some well-known ICU scoring systems, i.e. SAPS II [1] and APACHE III [2], chart data reflect physiological conditions of the patients. Thus, we use the chart data as low-level numerical features to characterize patient conditions directly. Note data that is in text format consist of descriptive and commenting information from caregivers. Data preprocessing operations are required to remove noise and ambiguity. For instance, some misspellings and abbreviations have to be processed. Besides, inconsistent recordings are required to unify due to different metrics, i.e. body temperatures in Celsius or Fahrenheit. Because note data is unstructured, it is very difficult to be applied directly to most of the existing machine learning algorithms. In this paper, we use a latent variable model, i.e. LDA, to extract topic distributions in medical notes as descriptive features for each note. After the two types of features have been extracted, we propose to use the Bag-of-Words (BoW) model to encode features. Features with different time stamps are incorporated into a unique representation. In this way, we get the feature representation for each patient.

We build our work based on the hypothesis that if the most discriminative features for each disease could be used in the learning process, we can get a more effective approach to exploit the disease data. In this paper, we propose a novel approach to learn from the disease data with two intuitive steps. Firstly, for each disease, we perform clustering on both positive and negative instances, and then construct the features specific to the disease by querying the clustering results. Secondly, we induce a family of classifiers with each of them being derived from the generated disease-specific features other than the original features.

It is worthwhile to highlight the following aspects of the proposed approach here.

1. Two different levels of features are extracted from structured and unstructured data. After that, Bag of Words (BoW) model is applied to encode features for representing health record of each patient.
2. For each disease, we construct a disease-specific feature so as to benefit the discrimination of different diseases. Besides, the proposed approach is scalable.
3. Extensive experiments are conducted on a real-world ICU patient dataset to evaluate the proposed approach. The experimental results have shown that our approach beats other alternatives, which confirms the effectiveness of the proposed approach.

The rest of this paper is organized as follows: Section 4 reviews related work on feature encoding for medical data and multi-label classification. We introduce the database, data preprocessing and the proposed algorithm in Section 2. Extensive experiments are conducted in Section 3. Section 5 concludes this paper.

2. Methodology

In this section, we first introduce the details of the used database, following by data pre-processing methods that used in this paper.

2.1. Database and data pre-processing

Multiparameter intelligent monitoring in intensive care II (MIMIC II) [3] is a real-world medical database that is publicly available. Over seven years (from 2001 to 2007), the database has successfully accumulate 32,535 patients' records in Intensive Care Units at Boston's Beth Israel Deaconess Medical Center. To protect privacy, experts remove all the Protected Health Information (PHI) and adopt more methods to prevent intentionally locating a specific patient before releasing the database. For example, all names in the records are replaced with deidentified notations, i.e. *pt*, all dates in a record are uniquely and randomly shifted into the future. To the best of our knowledge, MIMIC II is the largest and most complete ICU database that is publicly published in the world. In this database, there are different types of data that have been monitored and stored: (1) data recorded from bedside monitors for patients, e.g. waveforms and trends; (2) data from the clinical information system; 3) data from hospital electronic archives; (4) mortality information from the Social Security Death Index (SSDI). In this paper, we have used two major data sources in MIMIC II: medical chart event data and medical note data. The chart data are taken by the caregivers from medical device recordings, e.g. bedside monitors. Thus, it can reflect health conditions of a patient at a low level. In contrast, medical notes are taken by the professional staff in hospital, such as