Journal of Biomedical Informatics 84 (2018) 1–10

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Prediction task guided representation learning of medical codes in EHR

Liwen Cui[a], Xiaolei Xie[a,*], Zuojun Shen[a,b]

[a] Department of Industrial Engineering, Tsinghua University, Beijing, China
[b] Department of Industrial Engineering and Operations Research, University of California Berkeley, Berkeley, CA, USA

ARTICLE INFO

Keywords:
Representation learning
Word embedding
Natural language processing
Medical code
Electronic health records
Healthcare resource utilization

ABSTRACT

There have been rapidly growing applications using machine learning models for predictive analytics in Electronic Health Records (EHR) to improve the quality of hospital services and the efficiency of healthcare resource utilization. A fundamental and crucial step in developing such models is to convert medical codes in EHR to feature vectors. These medical codes are used to represent diagnoses or procedures. Their vector representations have a tremendous impact on the performance of machine learning models. Recently, some researchers have utilized representation learning methods from Natural Language Processing (NLP) to learn vector representations of medical codes. However, most previous approaches are unsupervised, i.e. the generation of medical code vectors is independent from prediction tasks. Thus, the obtained feature vectors may be inappropriate for a specific prediction task. Moreover, unsupervised methods often require a lot of samples to obtain reliable results, but most practical problems have very limited patient samples. In this paper, we develop a new method called *Prediction Task Guided Health Record Aggregation* (PTGHRA), which aggregates health records guided by prediction tasks, to construct training corpus for various representation learning models. Compared with unsupervised approaches, representation learning models integrated with PTGHRA yield a significant improvement in predictive capability of generated medical code vectors, especially for limited training samples.

1. Introduction

The adoption of the health information technology significantly increases the availability of digital healthcare records [1], such as Electronic Health Records (EHR) and insurance claims data. The EHR dataset used in this paper contains all the inpatient records over three years from five major hospitals in Beijing, China, with approximately 750,000 cases. In recent years, a large number of researchers have utilized machine learning techniques to uncover hidden information from complex healthcare data structures [2–4], which has potential to improve the quality of hospital services and achieve operational excellence. The raw EHR cannot be used directly in healthcare analytics. A fundamental and important step in model development is to convert medical codes in EHR to feature vectors. These medical codes are heterogeneous, indicating the diagnoses or procedures of the patients.

Since the representations of clinical information can significantly impact the performance of prediction models, efficient representation learning of medical codes is crucial for EHR analysis. Most previous approaches are unsupervised [5–7], which provide general purpose feature vectors that may perform poorly on a specific prediction task. Moreover, compared with supervised learning, unsupervised learning typically requires much more samples to achieve meaningful results [8], while in practice, the number of available patient samples is often relatively small. Therefore, the enhancement of the predictive capability of medical code representations given limited patient samples has significant practical implication.

The objective of this paper is to develop a representation learning model, which can learn vector representations of medical codes that have strong predictive capability for various prediction tasks, and requires relatively small amounts of training data.

1.1. Background

The choice of data representations has a large impact on the performance of machine learning models based on it [9]. There has been growing interest in representation learning, which aims to learn satisfactory data representations that are useful as input to prediction models, and has been successfully used in speech recognition [10], signal processing [11], object recognition [12], Natural Language Processing (NLP) [13], etc. For the analysis of EHR, finding appropriate vector representations for medical codes is one of the key challenges [14]. This task is challenging for the high dimensionality and event

[*] Corresponding author.
E-mail address: xxie@tsinghua.edu.cn (X. Xie).

https://doi.org/10.1016/j.jbi.2018.06.013
Received 14 February 2018; Received in revised form 14 June 2018; Accepted 16 June 2018
Available online 19 June 2018
1532-0464/

sparsity of medical codes. In other words, we usually need to face tens of thousands of distinct medical codes, while most of them only appear for a few times even in a very large dataset.

1.1.1. Word embedding from NLP

In NLP, a similar problem with the same challenges is how to represent words in large samples of language data. Representation learning techniques that map words to vectors of real numbers is referred to as *word embedding*. Various word embedding methods have been developed during the past decades, such as Latent Semantic Analysis (LSA) [15], tensor decomposition [16], and models from neural network, such as Skip-gram and Continuous Bag-of-words (CBOW) [17,18]. These word embedding models map words to a low-dimensional continuous space, where words closer in semantic meaning will have vectors closer to each other in this space. The semantic similarity between words is measured from statistical perspective. A corpus consists of sentences is required for the training of word embedding models. These sentences provide contexts for each word. Two words that have similar contexts when scanning through the corpus are considered to be semantically similar [19,20].

1.1.2. Word embedding used in EHR analysis

Recently, some researchers have utilized word embedding methods to learn vector representations of diagnosis and procedure information in EHR, and obtained promising performance in various clinical prediction tasks. Ho et al. [5] use tensor decomposition to map raw EHR data into meaningful medical concepts, and evaluate the result by a classification task of predicting high cost beneficiaries. Choi et al. [6] use Skip-gram to learn low-dimensional vector representations of diseases, medications, and procedures in longitudinal EHR. The obtained representation significantly improves the classification accuracy for onset of heart failure. Tran et al. [7] use restricted Boltzmann machine to map medical objects in EHR to low-dimensional vector space. The derived representation has satisfactory performance on suicide risk assessment for mental health patients.

1.1.3. Prediction task guided representation learning

Yang et al. [14] propose an approach that learns diagnosis and procedure information in EHR with supervised tensor decomposition, which solves an optimization problem with both representation learning and prediction modeling in the objective. Compared with unsupervised tensor decomposition, this approach significantly improves the predictive capability of the generated feature vectors. However, this approach is tightly integrated with tensor decomposition that stems from multilinear algebra, and hard to generalize to other representation learning models, such as Skip-gram, which can learn high-level factors that are related in highly nonlinear ways and beats the state-of-the-art in many NLP tasks [9].

The aim of the work described in this paper is to learn low-dimensional vector representations of medical codes in EHR, that have good predictive capability for various prediction tasks. We propose a novel approach called *Prediction Task Guided Health Record Aggregation* (PTGHRA). PTGHRA constructs training corpus for word embedding models by aggregating health records guided by prediction tasks, while most previous research uses one health record as a medical sentence. A brief comparison of these two treatments is shown in Fig. 1. PTGHRA can be easily integrated with various word embedding models. Compared with unsupervised approaches, PTGHRA significantly improves the predictive capability of obtained feature vectors, especially for limited training samples. We show in numerical experiments that for training set sizes smaller than 20,000 records, PTGHRA achieves up to 32% accuracy improvement in the prediction of healthcare resource utilization.

The rest of this paper is structured as follows. Section 2 describes PTGHRA in detail, and introduces the dataset we use. Section 3 shows the effectiveness of PTGHRA. In Section 4, several training corpus construction schemes are compared, the performance of PTGHRA on different prediction tasks and dataset sizes are provided, the limitations of this work are introduced. In Section 5, we conclude this paper and provide proposals for future research.

2. Materials and methods

2.1. Model overview

The main process of this research is illustrated by Fig. 2. In the EHR dataset, each health record contains medical codes, and other information such as age, gender, the number of hospitalizations, etc. We use representation learning methods to map medical codes to continuous vectors. Next, medical code vectors are combined with other information in health record to form feature vectors for prediction tasks. Specifically, we focus on the prediction of cost and Length of Stay (LoS), which are both common used measures of the utilization of healthcare resources [21,22].

The representation learning process is decomposed into two stages: medical sentence construction and word embedding model training. Here, the “medical sentence” is artificial compositions of medical codes, which provides contexts for each medical code. It is not clinical narrative notes, which are not contained in our EHR dataset. Word embedding models are trained with the corpus that is composed of these medical sentences, and generate medical code vectors. As we have mentioned in Section 1, medical codes with similar contexts will be mapped to low-dimensional vectors that are close to each other. PTGHRA is implemented on the medical sentence construction stage. This method decides the contexts of each medical code, and thus influences medical code vectors generated by word embedding models. Specifically, PTGHRA constructs medical sentences by aggregating similar health records guided by prediction tasks. Thus, the information of prediction tasks is contained in contexts of medical codes, which is supposed to have positive influence on the predictive capability of generated medical code vectors.

In this paper, PTGHRA is guided by healthcare resource utilization. The performance of PTGHRA is evaluated by the predictive capability of the generated vector representation of medical codes, i.e. the prediction accuracy of the regression model that is established based on it.

2.2. Dataset

Our EHR data are extracted from Hospital Quality Monitoring System (HQMS), which is a national level database developed by Ministry of Health of China. This dataset contains all the inpatient records with approximately 750,000 cases, from five major hospitals in Beijing, China, from January 2013 to December 2015. Each record contains demographic information (including age and gender), the utilization of healthcare resources (including LoS and cost), medical codes (including diagnosis codes and procedure codes), and other information (including insurance type, inpatient department, number of hospitalizations, etc.). In our EHR dataset, diagnoses are coded following a refined version of diagnostic ICD-10 codes, which adds one to two characters at the tail of the original ICD-10 codes [23], to express more clinical details without disrupting the hierarchical structure of the ICD coding rule. Similarly, procedure codes are coded following a refined version of procedural ICD-9-CM codes, which adds three characters at the tail of the original ICD-9-CM codes [24]. The medical codes related to a patient can be further categorized as primary codes and secondary codes, where primary codes are the main causes for the stay in hospital, or expend most of the utilized healthcare resources. Each patient has one primary diagnosis code, at most one primary procedure code, at most ten secondary diagnosis codes, and at most nine secondary procedure codes. In this research, we use medical codes, age, gender, and number of hospitalizations to predict LoS and cost jointly. We provide five data samples in Table 1.