Medical Concept Embedding with Variable Temporal Scopes for Patient Similarity

Zhihuang Lin, Dan Yang

***Abstract*—Electronic Health Records (EHRs) provide the possibilities to improve patient care and promote clinical research. In recent years, there has been an exponential increase in the range of diseases, diagnostic tests, and treatment regimens, which complicates the decision-making processes of doctors. Therefore, evaluating the clinical similarity of patients can provide effective and timely treatments and diagnoses for patients, which can allow doctors to make better decisions in a shorter time and at lower cost. In particular, the traditional machine learning methods for patient similarity are difficult to utilize the temporal information effectively while the temporal information in EHR data is very useful. In this paper, we propose a novel framework, called Patient Similarity Evaluation (PSE). Specially, PSE incorporates the temporal information to medical concept embedding for the representation learning of patients. Furthermore, PSE combines Siamese Convolutional Neural Network (CNN) with Spatial Pyramid Pooling (SPP) to measure the similarity between all patient pairs, which can predict the future health status of patients in advance and with precision. Experimental results demonstrate that our proposed framework outperforms all baseline methods.**

***Index Terms*—Patient Similarity; Medical Concept Embedding; Temporal Information; Siamese CNN with SPP**

I. Introduction

With the tremendous growth of the adoption of EHRs, a wealth of healthcare information including medication, procedure and diagnosis data are important resources for biomedical researchers to develop quantitative models for identifying similar patients. Through deep mining and analysis of EHR data, the doctors can find similar patients, which will enable to improve the probability of successfully curing patients a lot.

A. Motivation

Patient similarity studies [1] may reveal how potential clinical decisions would affect the development of patients' conditions. Patient similarity aims to derive a meaningful distance metric in the clinical field to measure the relative similarity among patients according to their health records. EHRs are the most fundamental, primary and meaningful basic data in the medical healthcare domain. The comprehensive analysis of EHR data will help doctors to judge patients' problems more accurately so as to provide patients with preventive and rehabilitative advice in advance. Therefore, EHRs are an available resource for measuring the clinical similarity between all patient pairs. In addition, the analysis of patient similarity is based on universal distance assessment between patients, and obtains the general rules of disease development from a large number of clinical practice data, which provides the possibility for computer-aided clinical decision support and personalized diagnosis and treatment using a general framework. Consequently, how to accurately and precisely measure patient similarity is an important and challenge issue.

Manuscript received March 18, 2020; revised June 23, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant No. 6167214, in part by General Scientific Research Projects of Liaoning Province under Grant No. 2019LNJC07, and in part by University of Science and Technology Liaoning Talent Project under Grant No. 601011507-22.

Zhihuang Lin, is with School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan, China (e-mail: lzh_ustl@ 163.com).

Dan Yang, the corresponding author, is a professor with School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan, China (e-mail: asyangdan@ 163.com).

B. Challenge

The effective representations for medical concepts (e.g., demographics, diagnostic history, medications, procedures, laboratory test results) of patients is crucial to patient similarity learning. A proper similarity measure enables various downstream applications, such as risk factor identification [2], medical diagnoses [3-4], morality prediction task [5] and clinical knowledge extraction [6]. However, most of the existing methods on learning effective vector representations of medical concepts for deriving the patient similarity measures still face many challenges:

1) Non-regularity: The patient information from EHRs is scattered and irregular. Therefore, we need to extract structured knowledge from EHRs for obtaining the valuable medical knowledge and experience.
2) Temporality: The process of patient treatments varies over time. As a result, medical concepts will change over time. By taking into account the temporal information of medical concepts, we can learn the effective representations of patients.
3) High-dimensionality: EHR data includes past and present medical records. Each record of EHRs, collected for a specific patient, consists of diagnoses, medication orders, laboratory test results and physiological parameters. Therefore, EHR data is usually represented in a high dimensional space.

C. Solution

Taking into account all challenges mentioned above, inspired by the idea of Word2Vec [7-9], we propose a novel framework to represent medical concepts of patients as the fixed-length vectors and derive a similarity measure between all patient pairs based on it. One way to represent medical concepts is one-hot vectors. However, the one-hot vectors have a high dimension and cannot reflect the semantic

relationships between medical concepts. Another way is Continuous Bag-of-Words model (CBOW) and Skip-gram model [7]. Though the two models can reflect the semantic relationships between medical concepts, they only consider the co-occurrences of medical concepts within a fixed-size window as indications of contexts, which may ignore the temporal characteristic in EHRs. Hence, the key of medical concept embedding is how to represent medical concepts effectively without loss of temporal information. There are two parts in our proposed framework: representation learning of patients and patient similarity learning. In representation learning of patients, we extend the Skip-gram model by adopting the variable temporal scopes in order to learn the effective representations of medical concepts, which takes the temporal information of EHRs into account. Based on the learned embeddings of medical concepts, we stack all medical concepts' embedding vectors in the medical history of patients to obtain the effective patient representations which are embedding matrices. In this way, the feature matrices of patients preserve the temporal properties in EHRs and mirrors the semantic relationships among medical concepts. In patient similarity learning, we deploy a patient similarity matching method based on Siamese CNN [10-11] to compute the similarity score between all patient pairs. After obtaining the similarity information indicating the risk level of the patient pair developing the same disease, we can better support clinical data mining work and acquire knowledge from retrospective data, thus supporting retrospective clinical research and achieving the goal of continuous improvement and improvement of medical quality.

D. Contributions

The main distinctive technical contributions of our work are summarized as follows:

1) We analyze the challenges of patient similarity learning and extend the Skip-gram model by leveraging the variable temporal scopes. The model converts patients' medical concepts to the fixed-length vectors which can preserve the semantic information between medical concepts and temporal information of EHRs at the same time.
2) We incorporate Siamese CNN with SPP as a deep learning model to measure the similarity between all patient pairs. The model can deal with the patient matrices of arbitrary sizes. To our best knowledge, Siamese CNN with SPP is the first to measure the similarity between all patient pairs.
3) We conduct extensive experiments on the large real dataset MIMIC-III, which significantly demonstrates that our proposed framework outperforms four baseline methods in terms of hospital readmission rate and incident rate difference for mortality. Moreover, comparative experiments are conducted between our proposed framework and the state-of-the-art methods on disease cohort classification and patient clustering, and our experimental results demonstrate that our proposed framework has the best performance.

The rest of this paper is organized as follows. Section 2 introduces the related work on patient similarity and medical concept embedding. We discuss our proposed framework in Section 3. In Section 4, we conduct experiments to compare our proposed framework with all baseline methods. Section 5 concludes the paper and our future work.

II. Related Work

In recent years, there are an increasing number of studies on evaluating the clinical patient similarity and representation learning in the medical healthcare domain. We firstly have a brief review in terms of patient similarity, and then review some related work on representation learning in the medical healthcare domain.

A. Patient Similarity

There are a lot of works concentrating on patient similarity in the field of health informatics. For example, Reference [12] deployed a cosine-similarity-based patient similarity metric (PSM) to weight the patient similarity measures. Reference [13] used the Tanimoto Coefficient (TC) to compute similarities between all patient pairs. Reference [14] proposed a locally supervised metric learning which is used for measuring similarities between patients represented by multi-dimensional time series. In [15], Wang applied the Triplet architecture to study fine-grained similarities among patients, which is used for fine-grained image similarity learning. Nguyen *et al.* [16] proposed the sequential matching procedure to calculate the distance between two patients, which can utilize the sequential order of medical concepts. However, these patient similarity matching methods do not take into account the temporal information in EHRs. Therefore, Wang *et al.* [17] presented a convolutional matrix factorization for detection of temporal patterns, and Cheng *et al.* [18-19] proposed an adjustable temporal fusion scheme using CNN extracted features. Reference [20] proposed Integrated Method for Personalized Modelling (IMPM) to provide personalized treatments and personalized drug designs.

B. Representation Learning in the Medical Healthcare Domain

Great progress has been made in Spoken Language Processing, Natural Language Processing (NLP) and Image Target Recognition. Many researchers utilize representation learning in the medical healthcare domain, for the reason that the sequence of medical codes can be seen as a natural language text. In recent years, many researchers have applied representation learning in the medical healthcare domain because an effective feature representation can simplify the difficulty of dealing with a problem and provide convenience for further applications. De Vine *et al.* [21] learned the representations of UMLS concepts from free-text patient records and medical journal abstracts. The Med2Vec model proposed by Edward Choi *et al.* [22] is a multi-layer representation learning tool for learning the representation of medical concepts and visit representations from EHR datasets. Youngduck Choi *et al.* [23] applied the Skip-gram model to

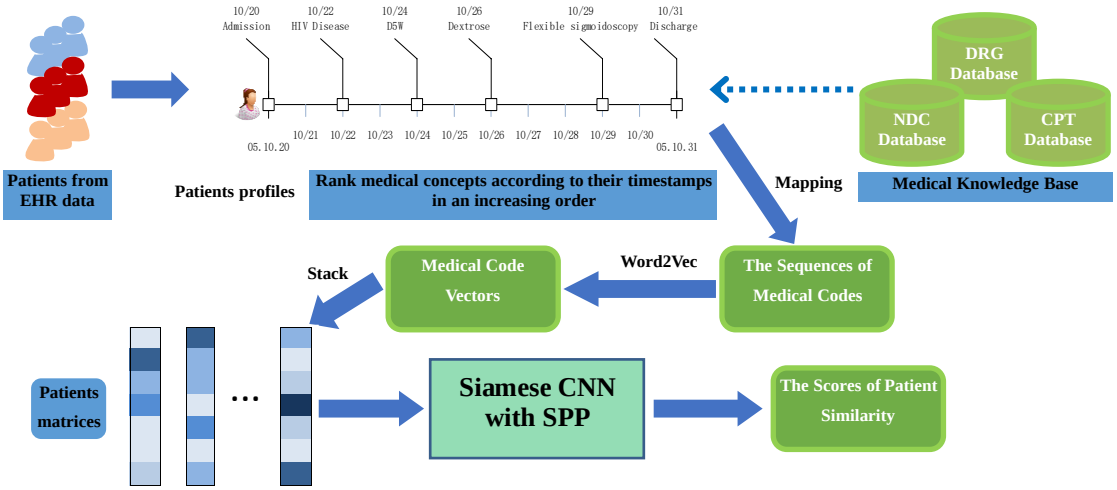


Fig. 1. We firstly rank the medical concepts according to their timestamps in an increasing order for each patient and map medical concepts to medical codes leveraging medical knowledge base (KB). Next, we employ Word2Vec with variable temporal scopes to medical code embedding. Then, we stack medical codes vectors of patients to obtain the patient matrices. Finally, we use the patient matrices as the input of Siamese CNN with SPP to compute the similarity.

medical concept embedding from medical journals, medical claims and clinical narratives, which is more useful than learning the embeddings from the clinical text. Meanwhile, some researchers employ the method with attention mechanism to learn the representation of medical concepts. Cai *et al.* [24] applied the attention mechanism to learn a "soft" time-aware context window for each medical concept in order to incorporate the temporal information to embed medical concepts. Mullenbach *et al.* [25] used an attentional CNN to encode the clinical text in EHRs, and then select different parts of the clinical text for predict diagnosis and treatments codes according to different labels.

III. The Proposed Framework

In this section, we introduce the details of our proposed framework on how to incorporate the temporal information of EHRs to medical concept embedding for representation learning of patients and measure the similarity between all patient pairs. The overview of our proposed framework is shown in Fig. 1. Our framework has four main phases. In phase 1, medical concepts of each patient are ranked in an increasing order according to their timestamps and each medical concept is mapped to a medical code leveraging medical KB (e.g., DRG database, NDC database and CPT database). Consequently, we can obtain medical code sequences which represent patients from EHR data. In phase 2, the sequence of medical codes is regarded as a natural language text describing a patient, and the extended Skip-gram model is applied to learn medical code vectors which represent the meaningful relations among medical codes. In phase 3, based on these medical code vectors, we construct the patient representation matrices which contain all medical features of patients. In phase 4, the patient matrices are fed into Siamese CNN with SPP to measure the similarity between all patient pairs.

A. The Sequence of Medical Codes

Our goal in this phase is to obtain medical code sequences which can effectively represent patients from EHR data. For each patient, by ranking all medical concepts in his/her EHR according to their happening timestamps (for medical concepts with the same timestamp we do not care about the order), we can obtain a medical concept sequence describing the historical condition of him/her. Moreover, each medical concept is mapped to a medical code leveraging medical KB. Finally, we can obtain medical code sequences for all patients. Given an ICU patient p whose associated medical concepts are ranked according to their timestamps in an increasing order, example of medical concept sequence for patient p mapping to medical code sequence is shown in Table I.

B. Medical Concept Embedding

After obtaining medical code sequences, we aim to learn the effective representations of medical codes by using the extended Skip-gram model which adopts variable temporal scopes. Compared with the one-hot encoding representation, our model can preserve the temporal information of EHRs and capture the latent relations among medical codes. In other words, medical codes which co-occur closely in time and have the similar scopes are mapped to the similar vectors so that their distance is small. Next, we first briefly review the Skip-gram model, and then describe how to incorporate the temporal information of EHRs to embed medical codes.

TABLE I
Example of Medical Concept Sequence Mapping

Patient	The Sequence of Medical Concepts	The Sequence of Medical Codes
p	Cranial Nerve Disorders, Coronary Bypass with Cardiac Catheter, Percutaneous Cardiac Procedure, Pioglitazone, Aspirin, Oxycodone-Acetaminophen, Acetaminophen, ...	73, 107, 1652, 64764045125.0, 17714001110.0, 406051262.0, 51079000220.0, ...

1) The Skip-Gram Model

The Skip-gram model in Word2Vec is applied to the medical healthcare domain mainly because the sequence of medical codes is treated as a natural language text describing a patient and medical codes are treated as words. The architecture of the Skip-gram model is shown in Fig. 2.

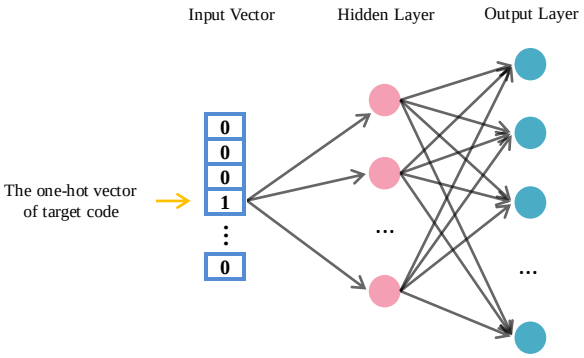


Fig. 2. The architecture of the Skip-gram model.

Formally, given a medical code sequence representing a patient $p = \{c_1, c_2, \ldots, c_N\}$, where N is the length of the medical code sequence, the Skip-gram model learns medical code vectors by using the target code c_t within a sliding window of size W to predict the context codes. It maximizes the log probability of predicting the context codes $c_{t-W}, \ldots, c_{t+W}$ which appear nearby the target code c_t:

$$\max \sum_{t=1}^{N} \log P(c_{t-W},\ldots,c_{t+W} \mid c_t) \tag{1}$$

Using an independence assumption, the probability in (1) is the following:

$$P(c_{t-W},\ldots,c_{t+W} \mid c_t) = \prod_{\substack{i=t-W \\ i\neq t}}^{i=t+W} P(c_i \mid c_t) \tag{2}$$

Therefore, Equation (1) is simplified to:

$$\max \sum_{t=1}^{N} \sum_{\substack{i=t-W \\ i\neq t}}^{i=t+W} \log P(c_i \mid c_t) \tag{3}$$

To reduce the computational complexity of (3), the Skip-gram model uses the Hierarchical Softmax [26] to approximate the probability distribution. Hierarchical Softmax aims to build a Huffman tree based on the frequency of each word in the vocabulary, which ensures the words with higher word frequencies are located at the leaf nodes in the shallow layer of the Huffman tree, and the words with lower word frequencies are located at the leaf nodes in the deeper layer of the Huffman tree. Furthermore, Hierarchical Softmax turns the problem into maximizing the probability of a specific path in the hierarchy (See Fig. 3.). If the path from the root to the medical code c_k is identified by a sequence of tree nodes ($b_0 = root, b_1, \ldots, b_{\lceil \log|N| \rceil} = c_k$), then

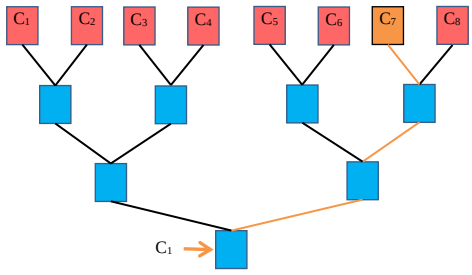


Fig. 3. Assume that the length of the medical code sequence for a patient is 8 ($N = 8$). Hierarchical Softmax factors out $P(c_7 \mid c_1)$ over sequences of probability distributions corresponding to the paths starting at the root and ending at c_7.

$$P(c_k \mid c_t) = \prod_{l=1}^{\lceil \log|N| \rceil} P(b_l \mid c_t) \tag{4}$$

Now, $P(b_l \mid c_t)$ is modeled by a binary classifier that is assigned to the parent of the node b_l as (5) shows,

$$P(b_l \mid c_t) = 1/(1+\exp(-v(c_t)\cdot\phi(b_l))) \tag{5}$$

where $\phi(b_l)$ is the representation vector assigned to tree node b_l's parent and has the same dimension as medical code vectors, and $v(c_t)$ is the representation vector of the target node c_t.

2) Variable Temporal Scopes

Usually, the Skip-gram model in Word2Vec predicts the surrounding contexts within a fixed-size sliding window by using a target word. But in the medical healthcare domain, the temporal information is significant for each medical concept because temporal information will reveal the relations between medical concepts. For example, diabetes mellitus typically lasts for several years, while dermatoses often lasts for several days, which indicates that a fixed-size context window will not work for all medical concepts. Therefore, we not only consider the chronological order of medical concepts, but also consider the context window size for each medical concept. We assume that the longer the medical concepts last, the larger the context window size of medical concepts. Therefore, we propose a method to determine the context window size of all medical concepts for each patient, which is inspired by the phenomenon that the medical concepts with a long duration appear more frequently and the medical concepts with a short duration appear less frequently. For each medical concept c_i of a patient p,

$$L(c_i,p) = \begin{cases} \alpha & f(c_i,p) < \alpha \\ f(c_i,p) & \alpha \le f(c_i,p) \le \beta \\ \beta & f(c_i,p) > \beta \end{cases} \quad (i = 1,2,\ldots,N) \tag{6}$$

where $L(c_i, p)$ is the context window size for medical concept c_i of patient p and $f(c_i, p)$ is the frequency of medical concept c_i in the EHR of patient p. α and β are the minimum and maximum size of context window respectively.

C. Effective Patient Representation

In the existing related works, usually a straightforward patient representation is constructed by converting all medical codes in his/her medical history to medical code vectors, then summing all those vectors to obtain a single representation vector. However, this patient representation will ignore the temporal information of EHRs. Therefore, we utilize a temporal representation: a patient is represented as an embedding matrix which has a dimension of $N_c \times d$, where N_c is the number of medical codes in the medical history of a patient and d is the dimension of all medical code vectors. Usually, N_c varies from patient to patient.

D. Patient Similarity Learning

We propose a deep learning model to measure the similarity between all patient pairs. The model is inspired by the text similarity problem tackled by Siamese LSTM Network [27]. Therefore, it is available to measure patient similarity using Siamese CNN. Each of the twin subnetworks of Siamese CNN uses this same CNN architecture. However, there is a technical issue in the training and testing of CNN: the fixed-size patient representations are taken as the input of CNN, which limits both the aspect ratio and the scale of the input. Consequently, in order to remove the fixed-size constraint of the CNN, we adapt the architecture of CNN by introducing Spatial Pyramid Pooling [28]. Siamese CNN with SPP maps the patient representation matrices of arbitrary sizes to the fixed-size vectors, and then computes the similarity score between the patient pairs.

In the following we will not only describe Siamese CNN and Siamese CNN with SPP in detail, but also explain how to compute the patient similarity score.

1) Siamese CNN

a) The Architecture of Siamese CNN

Siamese CNN combines Siamese Network [10] and CNN [11], and then maps the inputs to the target space and calculates the similarity in the target space by using a simple distance metric. Specifically, we train Siamese CNN to map the pair of temporal patient matrices to the fixed-size feature vectors respectively and then use the Cosine distance as the positive similarity function to express the degree of relatedness between the pair of patients. That is, the Cosine distance between the two patient vectors is taken as the final similarity score.

We assume that $X_1 = [v_1, v_2, \ldots, v_M]^T$ and $X_2 = [v_1', v_2', \ldots, v_M']^T$ are the temporal representation matrices of two patients p_1, p_2 respectively, where v_i and v_i' are the medical code vectors, and M is the length of medical code sequences. We use X_1 and X_2 as inputs to two identical CNNs with the same weights. Through the operation of two CNNs, we obtain the feature vectors $G_W(X_1)$ and $G_W(X_2)$. After obtaining the feature vectors of two patients, the similarity between the two patients is evaluated by the Cosine distance of two feature vectors. We will discuss how to construct the loss function in the following.

b) Loss Function of Siamese CNN

Suppose that the feature vectors of two temporal patient matrices X_1 and X_2 are G_1 and G_2 respectively. The similarity of two patients is the Cosine distance of their feature vectors, denoted $d(G_1, G_2)$. During the training phase of Siamese CNN, we use the contrastive loss function introduced by Chopra *et al*. in [29], which should satisfy the following two properties:

1) For two input patients of the same cohort, the greater the similarity, the smaller the loss function value.
2) For two input patients of different cohorts, the smaller the similarity, the smaller the loss function value.

The loss function of Siamese CNN is defined as shown in (7):

$$L(X_1, X_2, Y) = \frac{1}{2}(1-Y)d(G_1, G_2)^2 + \frac{1}{2}Y\max\{0, m - d(G_1, G_2)\}^2 \quad (7)$$

where the threshold $m > 0$ is a constant and Y is a binary label assigned to the pair of input patient matrices X_1 and X_2, so that $Y = 0$ indicates that the two input patients belong to the same cohort and $Y = 1$ indicates the opposite. When $Y = 0$, the second term of (7) is 0, and the first term of (7) becomes directly half the square of the Cosine distance of the two input patients. When $Y = 1$, the first term of (7) is 0, and the second term of (7) is the hinge loss. In the hinge loss, if the Cosine distance of two input patients is less than m, a penalty will be given. The greater the Cosine distance of two input patients, the smaller the penalty. If the Cosine distance of two input patients is greater than m, there will be no punishment.

2) Siamese CNN with SPP

In the existing CNN, the input size is generally fixed because the fully-connected layer of CNN requires a fixed number of neurons. Typically, the fixed-size input data is generated by a *crop* or *warp* operation. However, several problems will arise:

1) Forcing input data of arbitrary sizes to be converted to the fixed size may lose information.
2) *Crop* operation on input data may result in incompleteness, and *warp* operation may result in data deformation.

Therefore, when applied to the temporal patient representations of arbitrary sizes, we add an SPP layer behind the convolutional layer and before the fully-connected layer. This operation can address the constraint that the input size of

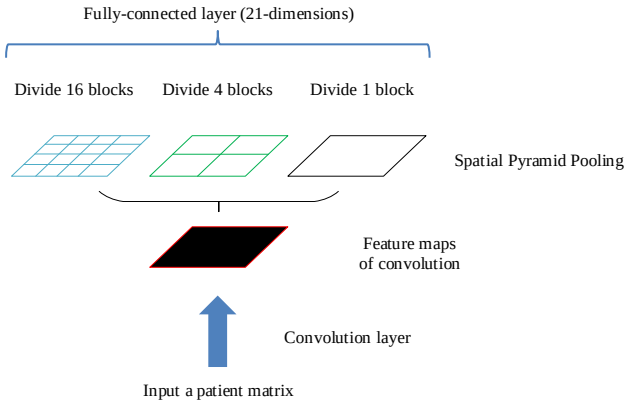


Fig. 4. The process of extracting features using SPP.

CNN is fixed. In the following we shall discuss how to extract features leveraging SPP.

As shown in Fig. 4, when we input a temporal patient matrix, the matrix is firstly entered into the convolution layer to get the feature maps, and then the feature maps are divided using scales of different sizes. Using three scales of different sizes (i.e., 4 × 4, 2 × 2, 1 × 1), we divide the feature maps into 21 (16 + 4 + 1) blocks. In the process of maximum pooling of spatial pyramids, the maximum value of each block is calculated separately in the 21 blocks, and then totally the 21 output neurons are obtained. Finally, the temporal patient matrix of arbitrary sizes is converted into a feature vector with 21 dimensions. For patient matrices of arbitrary sizes, after convolution and SPP layer processing, we can obtain the feature vector with the fixed dimension which is the number of neurons in the fully-connected layer.

It is worth noticing that the number of medical codes in the medical history of each patient is different. As a result, the size of representation matrix varies from patient to patient. Therefore, the spatial pyramid pooling strategy can be applied to Siamese CNN to extract spatial feature information for temporal patient matrices of arbitrary sizes. Fig. 5 presents the architecture of Siamese CNN with SPP.

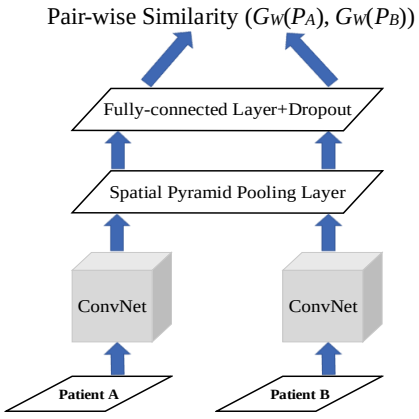


Fig. 5. The architecture of Siamese CNN with SPP.

We use two temporal patient matrices of different sizes as inputs to Siamese CNN with SPP. The two temporal patient matrices are respectively mapped into feature vectors of the same dimension through the two identical CNNs which have the SPP layer. In addition, we add the dropout layer to avoid overfitting. The main advantage of Siamese CNN with SPP is to overcome the defect caused by the temporal patient matrices of arbitrary sizes and improve the quality of patient clustering. Utilizing Siamese CNN with SPP makes the training data need not be normalized, and the effect is better than the traditional method.

E. Algorithm Description

Algorithm 1 represents our proposed framework for patient similarity learning—PSE in detail. The inputs of PSE are a set of patients, the minimum number of occurrences for medical codes, and the sequences of medical codes which have been ranked according to their timestamps in an increasing order. PSE has two main steps. Step 1: After mapping medical concepts to medical codes leveraging medical KB (Line 2), we train the Skip-gram model with variable temporal scopes from medical code sequences describing the patients to map each medical code into a fixed-length vector (Line 3-7). Step 2: For each patient from EHR data, the temporal patient representation is an embedding matrix which is constructed by stacking all medical codes vectors in his/her medical history (Line 8). However, some medical codes may not be in the tagged corpus, so the feature representations of these medical codes are the zero vectors. Afterwards, the patient matrices are fed into Siamese CNN with SPP to measure the similarity between all patient (Line 9). Additionally, for each patient, we select the patient corresponding to the highest similarity score (Line 10-16). Intuitively, since the patients have similar medical code sequences, it is highly possible that they have the risk of developing the same disease.

Algorithm 1 Medical Concept Embedding with Variable Temporal Scopes for Patient Similarity

Input: A set of patients $P = \{p_1, p_2, \ldots, p_N\}$; The minimum number of occurrences for medical codes λ; The sequences of medical codes $R = \{r_1, r_2, \ldots, r_N\}$.

Output: The most similar patient for each patient $p_i \in P$.

1: Configuration the total number of medical codes M, the number of medical codes for a patient K.
2: Map medical concepts → medical codes.
3: **foreach** $p_i \in P$ **do**
4: **foreach** $c_j \in r_i$ **do**
5: Select the medical codes with more than λ occurrences to construct the tagged corpus T.
6: Learn the length of context window for each medical concept according to the frequency of medical concepts $f(c_j, p_i)$.
7: Train the Skip-gram model to obtain the representation of medical codes $V = [v_1, v_2, \ldots, v_M]^T$.
8: Stack all medical codes vectors in the medical history of p_i to obtain the matrix $X_i = [v_1, v_2, \ldots, v_K]^T$.
9: Train the Siamese CNN with SPP using patient matrices as input.
10: $C \leftarrow \{\}$.
11: **foreach** $p_i \in P$ **do**
12: **foreach** $p_j \in P \setminus p_i$ **do**
13: Compute the similarity score between patient p_i and p_j.
14: Rank the similarity score.
15: Select the patient p_j corresponding to the highest similarity score.
16: $C \leftarrow p_j$.
17: Return C.

IV. Experiments

A. Dataset Overview and Preprocessing

Medical Information Mart for Intensive Care (MIMIC) III [30] is a database of intensive care patients opening to the public free of charge and collects data on ICU patients from Beth Israel Deaconess Medical Center between 2001 and 2012. The MIMIC-III dataset aims to develop and evaluate an advanced ICU patient monitoring system to improve the effectiveness, accuracy and timeliness of ICU clinical decision support. The MIMIC-III dataset consists of two parts, the clinical database and the physiological waveform database.

The clinical database has collected clinical information of

TABLE II
DATASET INFORMATION

Category	Descriptions	#Cardinality
Diagnosis	DRG codes	1,667
Medication	NDC codes	3,484
Procedure	CPT codes	2,018
Total		7,169

more than 60,000 ICU patients, including demographic characteristics of patients, discharge records, clinical and laboratory values, International Classification of Disease version 9 (ICD-9) codes associated with encounters, order, and referrals, procedure information in Current Procedural Terminology (CPT) [31] codes, diagnosis information in Diagnosis Related Groups (DRG) [32] codes and medication prescription information in National Drug Code (NDC) [33] codes and so on. Each record of ICU patients has detailed time information. The physiological waveform database records high-resolution waveform data from Philips bedside monitors such as electrocardiogram, blood pressure, pulse wave, and other physiological parameters such as respiration, blood oxygen, central venous pressure, etc. All of this data is subjected to rigorous de-identification processing. We choose the clinical database for our research. The patient features we used in our investigation are categorized into three groups as shown in Table II.

The first step of dataset preprocessing is to choose the patients which satisfy the following four conditions: (1) We remove the patients with missing data on admission date and discharge date; (2) We keep the patients which consist of at least thirty medical codes; (3) We remove the patients which have the discharge date after 2200/1/1; and (4) We remove the patients who have the missing data on diagnosis. Next, we choose the diagnosis information, medication information, and procedure information as the medical concepts of patients. Moreover, when we build the tagged corpus, we remove the medical concepts that are co-occurring less than three times (the medical concepts must have appeared in at least three medical concept sequences). Then, we map the selected medical concepts to medical codes leveraging medical KB and rank the medical codes according to their timestamps in an increasing order for each patient. Besides, we remove medication codes with missing data on start date and end date. Finally, we choose nine patient cohorts from the MIMIC-III dataset, namely, Atherosclerosis, Heart Failure, Kidney Failure, Intestinal Diseases, Liver Diseases, Pneumonia, Septicemia, Respiratory Failure and Gastritis. The remaining dataset contains 18,652 inpatient medical records, as shown in Table III.

B. Comparison Methods

To evaluate the effectiveness of the proposed PSE, we compare the framework with the following baselines and approaches in terms of different performance metrics.

1) PCA (Principal Component Analysis): A unsupervised method is widely used for dimension reduction and feature extraction [34]. We apply PCA on the one-hot EHR matrices of patients and perform Euclidean distance based on the PCA results.
2) PCM (Primary Code based Matching): A patient similarity method proposed by Lee *et al*. [35-36] that identifies patients who are most similar to each patient from EHR data. The method utilizes the first medical code between two medical code sequences for patient matching.
3) HDM (Hamming Distance based Matching): A method proposed by Hielscher *et al*. [37] that measures the patient similarity for complex objects contribute to class separation for a multifactorial disorder. The Hamming distance between two medical code sequences is the total number of medical codes where they mismatch.
4) CSM: Code Sum based Matching proposed by Choi *et al*. [38] obtains the patient representation by summing up all its medical code vectors, absolutely eliminating the sequential structure of medical codes. Firstly, CSM learns medical code vectors from EMRs using Word2Vec, a well-known embedding method. Then, it sums up medical code vectors of the patient to retrieve a single representation vector. Finally, the patient similarity score is the Cosine distance between their summed vectors.
5) Word2Vec-CNN: The method firstly learns medical code vectors from EHRs using the Skip-gram model in Word2Vec. Then we stack all medical code vectors in the medical history of patients to construct the patient matrices. Finally, the patient matrices pass through CNN to map into the feature vectors and we compute the similarity by the Cosine distance of the feature vectors.
6) T-Word2Vec-CNN: It applies the Skip-gram with

TABLE III
EXAMPLE FORMAT OF INPATIENT MEDICAL RECORDS

Subject ID	Hadm ID	The Sequence of Medical Codes	Disease Label
6	107064	302, 63739008901.0, 64253033335.0, 93008801.0, 472500360.0,...	KidneyFailure
13	143045	109, 0.0, 71015623.0, 45050130.0, 17714001110.0, 62584078833.0,...	Atherosclerosis
21	111970	7204, 416, 99254, 99291, 99291, 99253, 90935, 99291, 99231, 99254,...	Septicemia
109	175347	4603, 316, 59011010020.0, 59011010320.0, 781305714.0, 8084199.0,...	HeartFailure
111	192123	1394, 566, 173069502.0, 63739002401.0, 0.0, 0.0, 597007506.0,...	Pneumonia

variable temporal scopes to learn the embeddings of medical codes, which takes into account the temporal information of EHRs.

7) Word2Vec-Siamese: A Siamese network architecture instead of CNN to assess the similarity between the pair of patients.

C. Evaluation Metrics

With generated representation of each patient, we calculate the similarity score among all patient pairs using two different criteria: hospital readmission rate and incident rate difference for mortality. With the inherent difficulty of measuring the patient similarity, these two criteria are chosen since (1) both hospital readmission rate and incident rate difference for mortality play an significant role in many patient matching applications [35], [39] and (2) they are recorded in most routinely collected data, and hence have a broad prospect of application [40-41]. Furthermore, we evaluate the performance of patient clustering using two different criteria: Rand Index [42] and Normalized Mutual Information [43]. We will describe the detailed definition of these four criteria next.

1) The Hospital Readmission Rate (*HRR*)

Assume $P = \{p_1, p_2, \ldots, p_N\}$ is the collection of readmission statuses of N patients and $SP = \{p'_1, p'_2, \ldots, p'_N\}$ is the collection of readmission statuses of the most similar patients of N patients. *HRR* is computed as follows:

$$HRR = \sum_{i=1}^{N} \omega(P[i], SP[i]) / N \tag{8}$$

where $\omega(P[i], SP[i]) = \begin{cases} 0 & if\ P[i] \neq SP[i] \\ 1 & if\ P[i] = SP[i] \end{cases}$.

HRR measures the overall matching efficiency and $HRR \in [0, 1]$. In general, the greatest patient similarity has an *HRR* of 1 and the smallest patient similarity has *HRR* values close to 0.

2) Incidence Rate Difference for Mortality (*IRDM*)

Assume $P = \{(c_1, d_1), (c_2, d_2), \ldots, (c_N, d_N)\}$ is the collection of tuples (discharge date, death date) of N patients, where c_i is the discharge date, and d_i is the death date. The incidence rate of the collection of patients is computed as follows:

$$IR(P) = \frac{count(death)}{\sum_{i=1, d_i \neq null}^{N} (d_i - c_i) + \sum_{i=1, d_i = null}^{N} (d_{null} - c_i)} \tag{9}$$

where *count*(*death*) is the number of patients which have the death dates, d_i and c_i are death date and discharge date respectively, and d_{null} is 2200/1/1.

Similarly, we can compute the incidence rate of the most similar patients of N patients, called *IR*(*SP*). *IRDM* is computed as follows:

$$IR_{diff} = \left| IR(P) - IR(SP) \right| \tag{10}$$

IR_{diff} has lower bound of 0 corresponding to the perfect match between the partitions and upper bound of 1 that indicates the opposite.

3) Rand Index (RI)

RI is the most frequently used evaluation metric in data clustering. *RI* is computed as follows:

$$RI = (TP + TN) / \binom{n}{2} \tag{11}$$

where *TP* is the number of times a pair of patients belonging to the same cohort who are grouped into one single cluster. *TN* is the number of times a pair of patients from different cohorts who are grouped into different clusters. n is the total number of patients. In general, the larger the value of *RI*, the more consistent the clustering results are with the real situation.

4) Normalized Mutual Information (*NMI*)

NMI is often used in data clustering to measure the similarity of the two clustering results. *NMI* is computed as follows:

$$NMI(X, Y) = \frac{I(X, Y)}{\left[H(X) + H(Y)\right] / 2} \tag{12}$$

where Mutual Information $I(X, Y)$ is the relative entropy of the joint distribution $p(x, y)$ and the product distribution $p(x)(y)$, whose formula is:

$$I(X, Y) = \sum_{x} \sum_{y} p(x, y) \log \frac{p(x, y)}{p(x) p(y)} \tag{13}$$

$H(X)$ is the information entropy, and the formula is:

$$H(X) = -\sum_{i} p(x_i) \log p(x_i) \tag{14}$$

Similar to the value of *RI*, the closer the value of *NMI* is to 1, the better the quality of data clustering.

D. Parameter Setting

1) Medical Concept Embedding

For the Skip-gram model with variable temporal scopes, the sliding window size W is set to 30 and the learning rate is set as 0.05. We note that the minimum size of context window is the sliding window size (α = 30) and the maximum size of context window in our model is set as twice as the sliding window size (β = 60). The Skip-gram model with variable temporal scopes is trained with 500 epochs for the MIMIC-III dataset. The dimension of medical codes vectors d is set to 20, 50, 80, 100, 150, 200, respectively, for the comparison purpose, and after a serial of practices we select 100 as the dimension of medical codes vectors according to the best performance.

2) Patient Similarity Using Siamese CNN with SPP

In tensorflow, the parameters of Siamese CNN with SPP are as follows: the number of convolutional feature maps is set to 100. In the SPP layer, we use a 3-level pyramid. The

TABLE IV
TOP-3 MOST SIMILAR PATIENTS (PCM)

Patient (ID)	Nearest Patient (ID)	2nd Nearest Patient (ID)	3rd Nearest Patient (ID)
12359	336	1398	1541
340	48	49	347
7242	95	516	1013
606	48	49	347
2376	7	54	94

TABLE V
TOP-3 MOST SIMILAR PATIENTS (HDM)

Patient (ID)	Nearest Patient (ID)	2nd Nearest Patient (ID)	3rd Nearest Patient (ID)
12359	11319	2419	7269
340	4916	6797	10370
7242	7228	7632	10560
606	4916	1179	1569
2376	14438	4239	4514

TABLE VI
TOP-3 MOST SIMILAR PATIENTS (CSM)

Patient (ID)	Nearest Patient (ID)	2nd Nearest Patient (ID)	3rd Nearest Patient (ID)
12359	10873	9734	14130
340	2843	7581	7643
7242	13188	4735	7320
606	1748	2843	7190
2376	2334	2730	13104

TABLE VII
TOP-3 MOST SIMILAR PATIENTS (PSE)

Patient (ID)	Nearest Patient (ID)	2nd Nearest Patient (ID)	3rd Nearest Patient (ID)
12359	6198	10873	13955
340	321	7105	10581
7242	3868	2407	13188
606	15103	12704	2904
2376	2334	935	13104

pyramid is {4 × 4, 2 × 2, 1 × 1} (totally 21 bins). We use stochastic gradient descent [44] as the optimization method and contrastive loss as the loss function. We train Siamese CNN with SPP using 128 examples of shuffled mini-batches and adopt nonlinear rectification (ReLU) activation function. With regards to overfitting issue we add dropout regularization with dropout rate setting to 0.6.

E. Results and Analysis

1) Top-K Most Similar Patients

We run the proposed PSE and other three patient similarity learning methods to obtain the top-*k* most similar patients for each patient. We select 5 patients at random. Table IV, V, VI and VII describe the top-*k* ($k = 3$) most similar patients obtained by PCM, HDM, CSM and the proposed PSE, respectively. As shown in Table IV, V, VI and VII, the results of PCM and HDM are quite different from that of PSE. The main reason for this phenomenon is that PCM and HDM are the non-embedding methods for measuring patient similarity and the sequential information of patients from EHR data are not taken into account. For example, the top-3 most similar patients' IDs obtained by the patient ID 2376 using the PCM and HDM methods are completely different from the top-3 most similar patients' IDs obtained by PSE. However, the result of CSM is close to that of PSE because it is the embedding method for measuring patient similarity. As can be seen from Table VI and VII, the top-3 most similar patients' IDs obtained by the patient ID 2376 using the CSM method have something in common with that of PSE. This result might be due to the fact that they both utilize Word2Vec to learn the representation of medical concepts.

2) The Performance of Patient Similarity

In this section, we utilize the two criteria (*HRR* and *IRDM*) to evaluate the performance of patient similarity. We select 500 patients randomly and pick the most similar patient of each selected patient, and then use these two criteria to evaluate the performance of our proposed framework. Table VIII and IX are *HRR* and *IRDM* of the proposed PSE and other four patient similarity matching methods respectively. As can be seen from Table VIII and IX, the proposed PSE is obviously superior to other baseline methods for measuring the similarity between all patient pairs. The proposed PSE has the best performance in *HRR* and *IRDM*, which is 0.766 and 0.255, respectively. Comparing to the best performance, CSM achieves the second-best performance in *HRR* and *IRDM*, which is 0.684 and 0.336, respectively. The three patient similarity learning methods, namely PCA, PCM and HDM, achieve decline in *HRR* and *IRDM* compared with the other two methods for measuring the similarity between all patient

TABLE VIII
HOSPITAL READMISSION RATE (*HRR*)

Method	Technique	*HRR*
PCA	Principal Component Analysis	0.593
PCM	Primary Code Matching	0.614
HDM	Hamming Distance Metric	0.638
CSM	Word2Vec	0.684
PSE	Siamese CNN with SPP	0.766

TABLE IX
INCIDENCE RATE DIFFERENCE FOR MORALITY (1E-5)

Method	Technique	*IRDM*
PCA	Principal Component Analysis	0.420
PCM	Primary Code Matching	0.401
HDM	Hamming Distance Metric	0.384
CSM	Word2Vec	0.336
PSE	Siamese CNN with SPP	0.255

TABLE X
DISEASE COHORT CLASSIFICATION RESULTS

Method	Technique	Macro-AUC	Accuracy	Macro-F1
PCA	Principal Component Analysis	0.604	0.738	0.417
CSM	Word2Vec	0.726	0.792	0.446
PSE	Siamese CNN with SPP	0.818	0.891	0.534

pairs and PCA achieves the lowest performance in *HRR* and *IRDM*. This is probably due to the fact that PCA learns lower dimensional feature representations directly from the correlation matrix while not considering the semantic relationships among medical concepts. However, the semantic relationships can better reach the goal of producing meaningful representations of medical concepts. To sum up, context features are learned better in Word2Vec.

As we can see, *HRR* gains achieved by CSM are nearly 7% and 5% compared with PCM and HDM, respectively. CSM makes progress with nearly 7% and 6% in *IRDM* compared with PCM and HDM, respectively. CSM is superior to PSM and HDM mainly due to that CSM converts the medical concepts into the fixed-length vectors using the Skip-gram model in Word2Vec. Therefore, the method using words embedding has the better performance than the method without using words embedding. What is more, *HRR* gains achieved by the proposed PSE are nearly 15% and 13% compared with PCM and HDM, respectively. The proposed PSE makes progress with nearly 15% and 13% in *IRDM* compared with PCM and HDM, respectively. The proposed PSE achieves the most competitive performance because PSE not only takes into account the semantic relationships among medical concepts and the temporal information in EHR data, but also using a deep learning model to measure the similarity between all patient pairs. Overall, PSE achieves the best results on the large real dataset, demonstrating its generalizing ability in similarity learning of patients.

3) Disease Cohort Classification

We further investigate the effectiveness of the proposed PSE on disease cohort classification task. In the experiment, we successfully transform the patients of EHRs into the low-dimensional representations using different patient similarity learning methods including PCA, CSM and PSE, and apply MLP classification on the learned patient representations in order to correctly diagnose the diseases suffered by the patients. In addition, we use Macro Area Under The Curve (Macro-AUC), accuracy and Macro-F1 to evaluate the performance of disease cohort classification task, and use 10-fold cross-validation in which we randomly select 80% of the data for learning and the remaining 20% of data for testing the MLP classification.

Comparative results of different patient similarity learning methods for disease cohort classification task are shown in Table X. We observe that our proposed PSE achieves Macro-AUC of 0.818, accuracy of 0.891, and Macro-F1 of 0.534, which outperforms all the other methods, and CSM achieves the second highest performance. It is reasonable that the semantic information between medical concepts and temporal information of EHRs play the important roles in deriving meaningful information from EHRs. Thus, the embedding representations of patients obtained by PSE can enhance the performance of disease cohort classification. In general, our proposed PSE is a good choice in practice for disease cohort classification task due to its good performance.

4) Patient Clustering Results

We randomly choose 2,000 patients from 9 cohorts of diseases and use the two criteria *RI* and *NMI* to evaluate the performance of patient clustering. We adopt the proposed PSE and other three baseline methods to learn the representations of patients. Fig. 6 shows the results of two clustering criteria. We view that the proposed PSE achieves the best performance in *RI* and *NMI*, which are 0.785 and 0.727, respectively. T-Word2Vec-CNN is the second-best performance in the two criteria. Therefore, using the temporal information in EHR data makes the performance of patient clustering better than that of methods without using the temporal information. For the reason that we incorporate the temporal information in EHR data to medical concept embedding, the effective representations of medical concepts can be learned, which can make us obtain the better representations of patients. Another observation is that using Siamese CNN with SPP can achieve a higher score in the two criteria than the one using CNN model. The results indicate that using Siamese CNN with SPP can learn the better representation of patients from EHR data and achieve the better performance in measuring the similarity between all patient pairs.

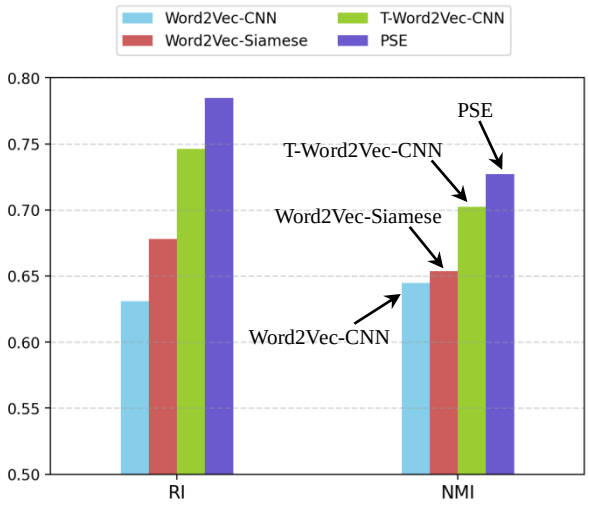
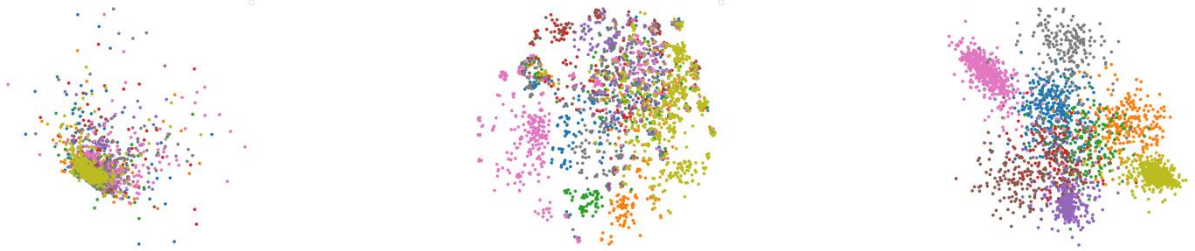


Fig. 6. Performance of Patient Clustering.



(a) PCA (b) CSM (c) PSE

Fig. 7. Visualization of patients. Each point indicates one patient. Color of a point indicates the cohort of the patient

5) Visualization

We utilize *t-SNE* [45] to reduce the dimension of medical code vectors, and plot the vectors of patients in a 2D space. T-SNE algorithm maps high-dimensional data to two or three dimensions suitable for human observation. We randomly choose 2,000 patients from 9 cohorts of diseases. As a result, each patient is mapped as a two-dimensional vector. Then we can visualize each vector as a point on a two-dimensional space. For patients which are labelled as different cohorts, we use different colors on the corresponding points. Therefore, a good visualization result is that the points of the same color are near from each other. The visualization figure is shown in Fig. 7.

From Fig. 7, we can see that the result of PCA is not satisfactory because the points belonging to different cohorts are mixed each other. For CSM, the clusters of different cohorts are formed. However, in the top part the patients of different cohorts are still mixed with each other and the boundaries of each group are not very clear. Obviously, the visualization of PSE performs best in both the aspects of group separation and boundary aspects.

F. Parameter Sensitivity of Medical Concept Embedding

The accuracy of the Skip-gram model depends on the parameters. In order to analyze the parameter sensitivity, we conduct the experiment in which we vary the dimension number of medical code vectors and report the results for the Skip-gram model.

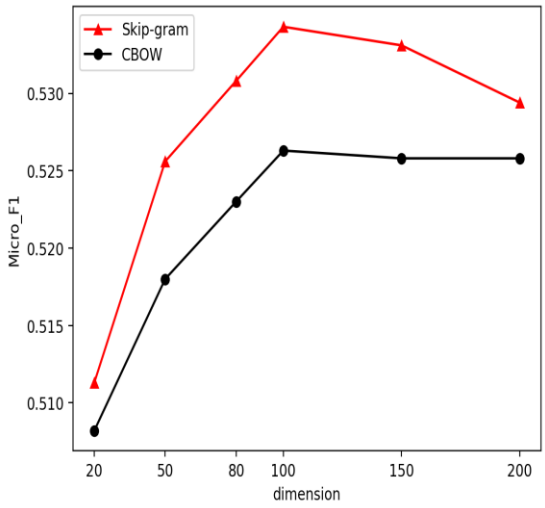


Fig. 8. Sensitivity of Word2Vec architectures with various dimensions.

Fig. 8 illustrates the Micro-F1 value (including training datasets and test datasets) of the Skip-gram model when using a specific number of dimensions. In the experiment, the Skip-gram model has the better performance than the CBOW model in medical concept embedding. The Skip-gram model performs better might be due to infrequent words (medical concepts) in the training corpus. On dimension $d = 100$ the result margin between both models is maximized and the Micro-F1 value reach its peak. On dimension $d < 100$ the Micro-F1 value between both models is increasing fast. On dimension $d > 100$ the Micro-F1 value between both models is decreasing slowly. We assume that this leads to some kind of overfitting and, thus, the optimal number of dimensions for medical concept embedding probably depends on the number of medical concepts and amount of training data.

In summary, we demonstrate that the Skip-gram model in Word2Vec performs best. It is interesting to see that the number of optimal dimensions for medical concept embedding must be geared to the underlying corpus.

V. Conclusions and Future Work

Due to the complexity of MIMIC-III dataset, extracting the effective representations of patients is vitally important. In the existing related works for medical concept embedding, many works usually overlook the temporal information in EHR data. In this paper, we present a patient similarity framework which exploits comprehensive semantic information among medical concepts and temporal information. The proposed PSE is divided into two parts. One is medical concept embedding for the representation learning of patients and the other is patient similarity learning leveraging Siamese CNN with SPP. The experimental results on the MIMIC-III dataset achieve the better performance compared with all baseline methods. Our next plan is to be going to mortality prediction task and other applications using our patient similarity framework.

References

[1] Sharafoddini, Anis , J. A. Dubin , and J. Lee . "Patient Similarity in Prediction Models Based on Health Data: A Scoping Review." Jmir Med Inform (2017).

[2] Ng, Kenney , et al. "Personalized Predictive Modeling and Risk Factor Identification using Patient Similarity." Amia Jt Summits Transl Sci Proc 2015(2015):132-136.

[3] Dongyang Li, Dan Yang, Jing Zhang, and Xuedong Zhang, "AR-ANN: Incorporating Association Rule Mining in Artificial Neural Network

for Thyroid Disease Knowledge Discovery and Diagnosis," IAENG International Journal of Computer Science, vol. 47, no.1, pp25-36, 2020.
[4] Riyanarto Sarno, Shoffi Izza Sabilla, Dedy Rahman Wijaya, and Hariyanto, "Electronic Nose for Detecting Multilevel Diabetes using Optimized Deep Neural Network," Engineering Letters, vol. 28, no.1, pp31-42, 2020.
[5] Allyn Jér?me, et al. "A Comparison of a Machine Learning Model with EuroSCORE II in Predicting Mortality after Elective Cardiac Surgery: A Decision Curve Analysis." PLOS ONE12.1(2017):e0169772-.
[6] M. Lamy, R. Pereira, J. C. Ferreira, F. Melo, and I. Velez, "Extracting clinical knowledge from electronic medical records," IAENG International Journal of Computer Science, vol. 45, no. 3, pp. 488–493, 2018.
[7] Mikolov, Tomas , et al. "Distributed Representations of Words and Phrases and their Compositionality." Advances in Neural Information Processing Systems (2013).
[8] Le, Quoc V. , and T. Mikolov . "Distributed Representations of Sentences and Documents." (2014).
[9] Bojanowski, Piotr , et al. "Enriching Word Vectors with Subword Information." Transactions of the Association for Computational Linguistics 5(2017):135-146.
[10] Bromley, Jane , et al. "Signature Verification Using a Siamese Time Delay Neural Network." Advances in Neural Information Processing Systems 6, [7th NIPS Conference, Denver, Colorado, USA, 1993] Morgan Kaufmann Publishers Inc. 1993.
[11] Krizhevsky, Alex , I. Sutskever , and G. Hinton . "ImageNet Classification with Deep Convolutional Neural Networks." Advances in neural information processing systems 25.2(2012).
[12] Joon, Lee , et al. "Personalized Mortality Prediction Driven by Electronic Medical Data and a Patient Similarity Metric." PLOS ONE 10.5(2015):e0127428-.
[13] Zhang, Ping , et al. "Towards Personalized Medicine: Leveraging Patient Similarity and Drug Similarity Analytics." Amia Jt Summits Transl Sci Proc (2014).
[14] Sun, Jimeng , et al. "Localized Supervised Metric Learning on Temporal Physiological Data." 20th International Conference on Pattern Recognition, ICPR 2010, Istanbul, Turkey, 23-26 August 2010 IEEE Computer Society, 2010.
[15] Wang, Yanda , et al. Learning Fine-Grained Patient Similarity with Dynamic Bayesian Network Embedded RNNs. Grundlagen des M&A-Geschäftes. 2019.
[16] Nguyen, Dang , et al. "Effective Identification of Similar Patients Through Sequential Matching over ICD Code Embedding." Journal of Medical Systems 42.5(2018).
[17] Wang, Fei , et al. "Towards heterogeneous temporal clinical event pattern discovery:a convolutional approach." Acm Sigkdd International Conference on Knowledge Discovery & Data Mining ACM, 2012.
[18] Cheng, Yu et al. "Risk Prediction with Electronic Health Records: A Deep Learning Approach." SDM (2016).
[19] Che, Zhengping , et al. "Exploiting Convolutional Neural Network for Risk Prediction with Medical Feature Embedding." (2017).
[20] Kasabov, Nikola, and Hu, Yingjie. "Integrated optimisation method for personalised modelling and case studies for medical decision support. " international journal of functional informatics & personalised medicine 3.3(2010):236-256.
[21] De Vine, Lance, et al. "Medical Semantic Similarity with a Neural Language Model." conference on information and knowledge management (2014): 1819-1822.
[22] Choi, Edward , et al. "Multi-layer Representation Learning for Medical Concepts." (2016).
[23] Choi, Youngduck , Y. I. Chiu , and D. Sontag . "Learning Low-Dimensional Representations of Medical Concepts." Amia Summits on Translational Science Proceedings 2016(2016):41-50.
[24] Cai, Xiangrui , et al. "Medical Concept Embedding with Time-Aware Attention." Twenty-Seventh International Joint Conference on Artificial Intelligence {IJCAI-18 2018.
[25] Mullenbach, James , et al. "Explainable Prediction of Medical Codes from Clinical Text." (2018).
[26] Mnih, Andriy , and G. E. Hinton . "A Scalable Hierarchical Distributed Language Model." Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008Curran Associates Inc. 2008.
[27] Shih, Chin Hong , et al. "Investigating Siamese LSTM networks for text categorization." 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) IEEE, 2017.
[28] Hadsell, R. , S. Chopra , and Y. Lecun . "Dimensionality Reduction by Learning an Invariant Mapping." Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on IEEE, 2006.
[29] He, Kaiming , et al. "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition." IEEE Transactions on Pattern Analysis & Machine Intelligence37.9(2014):1904-16.
[30] Johnson, Alistair Edward William , et al. "MIMIC-III, a freely accessible critical care database." Scientific Data3(2016):160035.
[31] Hameed, Haroon. "Current Procedural Terminology." (2017).
[32] Horn, and D. Susan . "Diagnosis Related Groups." Medical Care22.8(1984):777.
[33] Listed, Nos. "National drug code directory." 43.19(1969):80.
[34] Ablimit, Mijit , U. Yunus , and A. Hamdulla . "Signal Processing Teaching Case:a Tutorial on Principal Component Analysis." Modern Computer (2018).
[35] Joon, Lee , et al. "Personalized Mortality Prediction Driven by Electronic Medical Data and a Patient Similarity Metric." PLOS ONE 10.5(2015):e0127428-.
[36] Carnaby-Mann, Giselle D. , and M. A. Crary . "McNeill Dysphagia Therapy Program: A Case-Control Study." Archives of Physical Medicine & Rehabilitation 91.5(2010):0-749.
[37] Hielscher, Tommy , et al. "Using Participant Similarity for the Classification of Epidemiological Data on Hepatic Steatosis." IEEE International Symposium on Computer-based Medical Systems IEEE Computer Society, 2014.
[38] Choi, Edward , et al. "Medical Concept Representation Learning from Electronic Health Records and its Application on Heart Failure Prediction." (2016).
[39] "Information technology for healthcare transformation." IBM JOURNAL OF RESEARCH AND DEVELOPMENT 55.5(2011):6-6.
[40] Wang, Y. , et al. "Community-Level Association Between Home Health and Nursing Home Performance on Quality and Hospital 30-Day Readmissions for Medicare Patients." Home Health Care Management & Practice(2016):1084822316639032.
[41] H?Konsen, Sasja Jul , et al. "Nursing Minimum Data Sets for documenting nutritional care for adults in primary healthcare." JBI Database of Systematic Reviews and Implementation Reports 16.1(2018):117-139.
[42] Rand, William M. . "Objective Criteria for the Evaluation of Clustering Methods." Publications of the American Statistical Association 66.336(1971):846-850.
[43] Meil?, Marina . "Comparing clusterings — an information based distance." Journal of Multivariate Analysis 98.5(2007):873-895.
[44] Paras. "Stochastic Gradient Descent." Optimization (2014).
[45] Arora, Sanjeev , W. Hu , and P. K. Kothari . "An Analysis of the t-SNE Algorithm for Data Visualization." (2018).