Learning Graph-based Embedding from EHRs for Time-aware Patient Similarity

Hua Jiang, Dan Yang

Abstract—**With the wide application of electronic health records (EHRs), the research on mining effective medical knowledge from EHRs and supporting clinical decision-making has become an important research field, and patient similarity analysis is an important research direction. In this paper, we propose a time-aware patient similarity framework from EHRs, named T-PS. Specially, T-PS first constructs a high-quality temporal medical entity association graph from EHRs by converting the patient profile. The patient profile includes the diagnoses, medicines and procedures. Then the medical entities in the temporal medical entity association graph can be projected into a low-dimensional vector space. In the process of network representation learning, the time decay function is combined with the medical entity representations to obtain the temporal patients' representations. Finally, the patient similarity can be calculated by the cosine similarity among the patient representations. Experiments based on real-world ICU dataset MIMIC-III demonstrate the effectiveness and correctness of T-PS.**

Index Terms—**Patient Similarity, Time-Aware Information, Electronic Health Records, Network Representation Learning**

I. INTRODUCTION

With the continuous development of hospital health information systems and health websites, medical data such as medical activities, medical researches, and health information behaviors are increasingly abundant. Medical data are important resources to construct quantitative analysis model of patients. Patient similarity studies have been identified as one of the key technologies in medical reform. Patient similarity is based on the general distance evaluation between patients and obtains the general rules of diseases from a great deal of clinical practice data, which provides the possibility for computer-aided clinical decision support applications and personalized diagnosis and treatment using the general framework. Patient similarity has been applied to target patient retrieval [1], and clinical pathway analysis [3]. Although it is very significant, there are few types of research on patient similarity learning at present. Electronic Health Records (EHRs) [4] contain a large number of available medical data such as medications, diagnoses, procedures, lab results, etc. The medical data are diverse storage forms, sparse and high-dimensional, which have become the most important challenges in EHRs applications. Many studies have attempted to address the challenges inherent to EHRs. Therefore, deriving effective and robust representations for EHRs are a critical step for making various healthcare applications possible. In recent years, network representation learning [5-8] develops rapidly. Network embedding learns the node representation in the network, in which every node is mapped into low-dimensional vector space. Network embedding methods are further used for patient feature analysis and other applications [9-10]. Some studies show that the effect of embedding relies upon vector operations to represent learned word similarities [44]. Another study compounded medical embedding with human-selected features to enhance clinical representation [6]. Similarly, [45] designed a model for statistical script learning by using long short-term memory method. The model has been proved to work well in some artificial intelligence tasks. Though these studies have shown the improved performance on varies clinical tasks. To our best knowledge, there is no model that addresses all the mentioned questions earlier. Considering all the above challenges, we design a novel framework time-aware patient similarity base on network embedding from EHRs, named T-PS. This paper has three contributions, which can be summarized as follows:

1) We extract patient related information, including diagnoses, medicines and procedures from the EHRs and create temporal medical entity association graph to capture the associated medical entities for patients. Meanwhile, we use network embedding method, and fully consider the network structure information to learn the effective medical entity representations and patient representations. The patient representations can be used to calculate patient similarity.

2) We extract medical entities from EHRs and preserve temporal information. When we use different time interval medical entity, the medical entity and patient representations can as time change.

3) We evaluate the effectiveness of the T-PS on real-world ICU dataset MIMIC-III. Experimental results show that the performance of T-PS is better than other contrast methods.

The remaining sections of this paper are organized as follows. We introduce related works in Section 2. The preliminary is presented in Section 3. Details about our temporal medical entity association graph embedding framework in Section 4. Section 5 shows the experimental results. Finally, Section 6 summarizes the conclusions and future work.

II. RELATED WORK

A.Patient Similarity

In recent years, patient similarity as a fundamental problem has attracted great attention in the field of health

Manuscript received June 14, 2020; revised August 17, 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.

Hua Jiang, is with School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan, China (e-mail: jh_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).

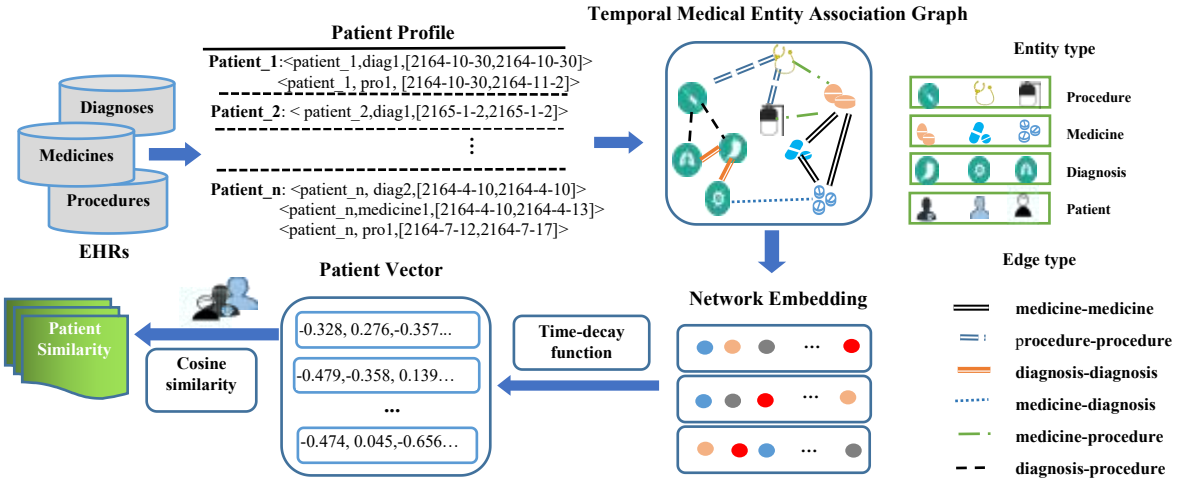


Fig. 1. T-PS: Framework of time-aware patient similarity from EHRs. We firstly construct the patient profile according to patient-related medical entity records for each patient. Next, we construct temporal medical entity association graph, $G_{T\text{-}MEA}$. Then, we leverage network embedding to obtain the medical entities vectors. At the same time, all the medical entities vectors are added according to time-decay function to get patients vectors. Finally, we use patients vectors as the input of cosine similarity to compute the similarity.

care. The patient similarity has the potential to aid clinical decision-making. For example, [1] used the generalized Mahalanobis measure method to calculate patient similarity. [2] proposed a novel algorithm based on support vector machine to measure the similarity. [3] mapped medical events to a low-dimensional vector space and extracted the patient representations to analyze the patient similarity. In [4], an adjustable time-decay fusion scheme based on CNN to extract features is proposed. [11] designed a personalized integration model to provide cures and medicines for similar patients. [24] proposed a convolution matrix decomposition approach to detect temporal patterns. [25] used k-means method to determine the similar patients' groups. [26] proposed a disease classification model based on three-layer deep learning method. There are many statistical and machine learning methods are proposed to analyze patient similarity. At present, may studies have proved that their methods are effective. However, the focus has primarily been on applying diagnosis data from EHRs for the learning task.

In this paper, we adopt a more holistic view of the patient and consider different sources of patient's information from EHRs, including diagnoses, medicines and procedures to develop a temporal medical entity association graph. We adopt network embedding to map medical entities into vectors and represent a patient by adding its associated medical entities. Finally, we use a patient similarity function to calculate patient similarity.

B.Network Representation Learning

Network representation learning is applied in many practical aspects. There are many researches focus on designing new embedding methods. Works in network embedding mainly consist of three categories: (1) Models based on matrix factorization; (2)Techniques based on random walk; (3) Deep learning methods enhance the ability of the model to gain non-linearity information in the network. According to whether the types of node and edge are the same, the information network can be divided into homogeneous network and heterogeneous network. Most of the existing homogeneous information networks use the existed depth models and combine the network features to learn the node representations and edge representations. LINE [12] tried to approximately factorize the adjacency matrix and captured first-proximity and second-proximity neighborhood nodes respectively to learn the node representations. DeepWalk [13] used truncated random walks to capture the context information for each node and utilized word2vec to learning node representations. Node2vec [14] explored different context nodes by using a biased random walks. In short, it can be regarded as an extension of DeepWalk, which combined DFS and BFS random walks. Heterogeneous information network representation learning has arisen and developed rapidly in recent years. Due to the complexity of content and structure information, heterogeneous information networks are difficult to obtain useful information. Metapath2vec [16] used meta-path random walks in heterogeneous networks to extract node structure information and used skip-gram [17] algorithm to learn node representations. HINE [18] calculated the similarity between nodes based on meta-path random walks. The model extracted the nonlinear features of the network structure using a deep automatic encoder. In addition to use the topology information of the network structure, there are also many methods to learn more accurate entity representation by using the content information or other auxiliary information of entities. HNE [19] extracted features from text and image data through CNN and MLP, and used transfer matrix to map different kinds of data into the low-dimensional vector space. As more and more node attribute information is observed and recorded in real life, how to extract useful information from network structure and several attributes information to learning a unified low-dimensional vector representation has become an important research topic. Attribute network representation learning arises at the historic moment. LANE [20] integrated tag information into attribute network representation learning, calculated similarity matrix between nodes, and took covariance as the measurement of matrix correlation. DANE [21] used a combination of an offline algorithm and an online algorithm to reduce the time required to learn dynamic attribute network representation. SNEA [22] solved the problem of symbol and node attribute

information fusion based on the structural balance theory in social psychology. In the framework of matrix factorization, TADW [23] integrated node text features into network representation learning.

III. Preliminaries

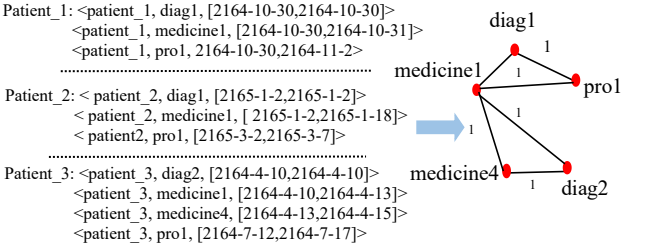
Before we focus on the patient similarity problem, we first give some definitions.

Definition 1 (Patient Related Medical Entity Record). EHR contains medical entities such as diagnoses, medicines and procedures, from which we extract these data to generate a patient related medical entity record. Each record is represented as a triple r_p =<p, e, t_s>, the triple means that patient p related medical entity e during time span t_s [t_b, t_e], where t_b *and* t_e are the begin time and end time of the time span t_s, respectively.

Definition 2 (Patient Profile). For patient p, the patient profile Dp is a set of patient p' related medical entity records, Dp={ r^1_p, r^2_p, ..., r^n_p } (n>1). These data are sorted by begin time t_b.

Definition 3 (Temporal Medical Entity Association Graph, $G_{T\text{-}MEA}$). $G_{T\text{-}MEA}$ is denoted as $G_{T\text{-}MEA}$= (V, E). V is a set of medical entities. We use v to represent each medical entity node. The set of medical entity type denoted as V_{type}. V_{type}={Diagnosis, Medicine, Procedure}. They represent medical entities diagnoses, medicines and procedures, respectively. E is a set of edges between medical entities, and e_{ij}(v_i, v_j) is an edge between node v_i and v_j. The set of edge type denoted as E_{type}. E_{type}={medicine−medicine, medicine−diagnosis, medicine−procedure, diagnosis−diagnosis, diagnosis−procedure, procedure−procedure}, they represent the co-relationships of medicines, diagnoses and procedures.

We use the patient profile to construct temporal medical entity association graph. Since each patient has multiple patient related medical entity records r_p in the patient profile, we combine the records of the patient in pairs. Given a time interval ΔT, for each patient related medical entity record pair{(r_p^i, t_b^i), (r_p^j, t_b^j)}, if 0<| t_b^i-t_b^j |<ΔT, the medical entities in the r_p^i and r_p^j co-occurring, and there is an edge between the two medical entities. The weight w_{ij} of edge represents the medical entities co-occurrences number within ΔT. For each patient in the patient profile, we find the co-occurrence of the medical entities according to the above description and finally form a temporal medical entity association graph. A simple example of generating temporal medical entity association graph from the patient profile is shown in Fig.2.



(a) Patient Profile (b) Temporal Medical Entity Association Graph, $G_{T\text{-}MEA}$

Fig. 2. An example of generating Temporal Medical Entity Association Graph, $G_{T\text{-}MEA}$ from the patient profile (ΔT=30 days). Notes: patient data is de-privatized.

With the above definitions, we can construct the temporal medical entity association graph, $G_{T\text{-}MEA}$ and obtain the medical entity representations. Because of the diagnoses given by doctors, the medicines are taken by patients, and the treatments received during hospitalization are dynamic and have a temporal relationship. The goal of constructing temporal medical entity association graph is to describe each patient with the corresponding medical entities.

IV. Temporal Medical Entity Association Graph Network Embedding

The next section will explain in detail how to learn medical entity representations and time-aware patient representations.

A. Temporal Medical Entity Association Graph Network Embedding

Given a temporal medical entity association graph, $G_{T\text{-}MEA}$. It is very important to make fully use of the latent information in $G_{T\text{-}MEA}$. According to the second-order proximity, we can assume that the more information shared between nodes, the more similar they will be. Network representation learning is to project the nodes' information and the relationship into a low-dimensional vector space. Each node v_i represents: 1) the node itself; and 2) the contexts of other nodes. For example, m_i, m'_i represent the node representations of the node v_i in two different roles, respectively.

For each edge e_{ij}(v_i,v_j) in the graph, we study how to define the probability of "context" v_j generated by node v_i. The probability can be calculated as follows:

$$p(v_j \mid v_i) = \frac{\exp(m_j'^{T} \cdot m_i')}{\sum_{k=1}^{|V|} \exp(m_j'^{T} \cdot m_i')} \tag{1}$$

In (1), $|V|$ is the number of context entity nodes, $P(.|v_i)$ represents the conditional probability of the context of all nodes v_i in the temporal medical entity association graph.

The empirical distribution $\hat{p}(v_j|v_i)$ is defined as:

$$\hat{P}(v_j \mid v_i) = \frac{w_{ij}}{sum_i} \tag{2}$$

where w_{ij} is the weight of the e_{ij}(v_i,v_j), sum_i= $\sum_{k \in V(i)} w_{ik}$ is the medical entity node v_i's out-degree summation, $V(i)$ is the v_i's neighbor nodes set.

We select KL-divergence [27] as a distance function to calculate the $P(.|v_i)$ and $\hat{p}(.|v_i)$ distance.

$$O_{G_{T-MEA}} = \sum_{v_i \in V} \lambda_i d(P(\cdot \mid v_i), \hat{P}(\cdot \mid v_i)) \tag{3}$$

where λ_i can be replaced by the degree of node v_i in the graph. In formula (3), some constants are omitted, and the objective function is as follows:

$$O_{G_{T\text{-}MEA}} = -\sum_{(v_i,v_j) \in E} w_{ij} \log P(v_j \mid v_i) \tag{4}$$

Due to the second-order similarity between the calculated nodes, the denominator calculation of the softmax function needs to traverse all nodes, which is very inefficient. The optimization is realized by using the technique of negative sampling [28]-[29]. We adopt negative sampling approach

to reduce computation complexity when computing formula(4). So the objective function can be calculated as formula (5):

$$O_{G_{\text{T-MEA}}} = -\sum_{(v_i,v_j\in E)} w_{ij}\{\log\sigma(m_j'^{T}\cdot m_i') + \sum_{k=1}^{K} E_{v_n\sim Z_n(v)}[\log\sigma(-m_j'^{T}\cdot m_i')]\} \quad (5)$$

where, $\sigma(x)$ is the sigmoid function. The value of noisy node distribution $Zn(v)$ is set to $Zn(v) \propto sum_v^{3/4}$, where sum_v is the entity node v's out-degree. K denotes the negative edges size.

Finally, we apply the asynchronous stochastic gradient algorithm (ASGD) to optimize formula (5).

B. Time-aware Patient Embedding

Since the diagnoses, medicines taken and treatments received by the patient during hospitalization are changing over time, recent medical entities related to the patient only indicate the patients' recent physical conditions. In the previous related work, many works did not consider the time information when carrying out patient representation learning. In order to catch the time features on patient embedding representation, we consider the time information when learning the patient embedding representation, and use the time-decay function to obtain the patient embedding representation.

Given time-aware patient vector representation $p'_{i,t}$, we can calculate $p'_{i,t}$ based on patient p_i before time t related medical entity representations. We can use formula (6) to obtain the time-aware patient representation:

$$p'_{i,t} = \sum_{((p_i,e_j,t_j)\in Dp_i)\cap(t_{b_j}<t)} \theta(t-t_{b_j})\cdot D'p_j \quad (6)$$

where D'_{p_j} is the vector representation of medical entity j. $\theta(\cdot)$ is some time decay function [30]-[32]. $\theta(\Delta t)$ increases as the Δt decreases. $\theta(\cdot)$ ensures that the medical entities appear later have larger weights.

We use medical entity representations and time decay function to achieve patient representations. The time decay function used in T-PS is as follows:

$$\theta(\Delta t) = \begin{cases} \frac{1}{2}[1+\cos(\frac{\Delta t\cdot\pi}{\sigma})] & if\ \ \Delta t \le \sigma \\ 0 & otherwise \end{cases} \quad (7)$$

C. Utilizing Patient Embedding for Patient Similarity

In this section, we will describe how to utilize patient vector representations to calculate patient similarity. Given an embedding representation of patient p_i at time t, we calculate patient's score against other patients. We use the formula (8) to measure the patient similarity:

$$score(p_i, p^q{}_j) = \frac{p'_{i,t}\cdot p'^{q}{}_j}{\left\|p'_{i,t}\right\|_2\left\|p'^{q}{}_i\right\|_2} \quad (8)$$

where $p'_{i,t}$ is the patient representation of $p_{i,t}$, and $p'^{q}{}_j$ represents the queried patient representation for patient $p^q{}_j$.

Once the embedding vector of the patient is obtained, we can obtain the patient of the former top-k nearest to the query patient according to the ranking.

V. Experiments and Evaluation

In this section, we will use the patient representations to analyze patient similarity. After obtaining the vector representations of the patients, we use t-SNE [33-34] to cluster and visualize the patient representations, and then evaluate three different network embedding methods. At the same time, we choose cosine similarity function as patient similarity function to calculate patient similarity.

A. Dataset

MIMIC-III [35] is real-world EHRs data from the ICU of Beth Israel Deaconess Medical Center. Meanwhile, it contains distinct 46,520 patients, 650,987 diagnoses and 1,527,702 prescription records that associated with 6,985 distinct diseases and 4,525 medicines. Each record of ICU patient has detailed time information. From dataset, we took out prescriptions, drgcodes and cptevents three tables which associated with patient's medicines, diagnoses and procedures, respectively. In these tables, different medical entities are represented using the International Classification of Diseases (ICD-9) [36], the Program Information Code (CPT) [37], the Diagnostic Related Information Group Code (DRG) [38], and the Drug Information Code (NDC) [39], respectively.

When using the dataset for experiments, preprocessing should be carried out first. If a patient with missing values, we will remove it. Then, we collect diagnoses, medicines, procedures of patients from EHRs to construct a patent profile. The patient profile is then used to build a temporal medical entity association graph.

B. Running Environment

The running environment is shown in Table I.

TABLE I
Running Environment

Parameters	Configuration
CPU	Intel(R) Xeon(R) E5-2620 v4 @ 2.10GHz
Memory Size	4G
Operating System	Windows 7
Development language	Python 3.6
Development framework	TensorFlow 1.14

C. Evaluation Metrics

Firstly, the patient representations are used to calculate the patient similarity. Then, we can obtain the most similar patients for the test patient. Exactly, we measure the patient similarity using four popular criteria: *SSE*, *Purity*, normalized mutual information (*NMI*) and hospital readmission rate (*HRR*). The four evaluation indicators are defined as follows:

1) Sum Of Squared Error (SSE)

SSE [40] is used for cluster analysis. It is defined as follows:

$$SSE = \sum_{i=1}^{n}\sum_{j=1}^{m} W^{(i,j)} = \left\|x^{(i)} - u^{(j)}\right\|_2^2 \quad (9)$$

where $u^{(j)}$ represents the center of the j cluster.

2) Purity

Purity is also used in data clustering. We compute (10) as defined in [41]:

$$Purity(X,Y) = \frac{1}{N}\sum_{k}\max_{j}\left|x_i \cap y_j\right| \quad (10)$$

Where $X=\{x_1, x_2, ..., x_k\}$ is the cluster partition, and $Y=\{y_1, y_2, ...,y_j\}$ is the real class partition. N is the total number of samples,.

3)Normalized Mutual Information (NMI)

NMI can be used as a measure of clustering similarity[42]. The range of *NMI* is [0,1]. The closer NMI is to 1, the more similar the cluster results are to the real data set. *NMI* can be defined as follows:

$$NMI(X,Y)=\frac{I(X,Y)}{[H(X)+H(Y)/2]} \tag{11}$$

where $I(X, Y)$ is mutual information, it can be calculated as follows:

$$I(X,Y)=\sum_x\sum_y p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{12}$$

where p(x, y) is joint distribution, p(x)(y) is product distribution.

$H(X)$ is the information entropy of divided objects, it is defined as follows:

$$H(X)=-\sum_i p(x_i)\log p(x_i) \tag{13}$$

4)The Hospital Readmission Rate (HRR)

We assume that $X=\{x_1,x_2,...x_p\}$ is the set of readmission statues of P patients and $Y=\{y'_{e1},y'_{e2},..., y'_{ep}\}$ is the set of readmission statues of the most similar patient of P patients. We calculate *HRR*[43] as follows:

$$HRR=\sum_{i=1}^{P}\omega(X[i],Y[i]) \tag{14}$$

$$\omega(X[i],Y[i])=\begin{cases}0, & X[i]\neq Y[i]\\ 1, & X[i]=Y[i]\end{cases} \tag{15}$$

HRR is used to measure overall consistency and $HRR\in[0, 1]$. Generally, the greatest patient similarity has an *HRR* of 1, and the smallest patient similarity has *HRR* close to 0.

D. Results and Discussion

1) Representation Learning Based On $G_{T\text{-}MEA}$

We use LINE, DeepWalk and node2vec to learn the representation of the temporal medical entity association graph. The parameters are set as follows.

- LINE: The epoch is 5000. The mini-batch size of the stochastic gradient descent is 1 for the network embedding method. The number of negative samples K is 5. Meanwhile, we set the time interval to 30 days and σ is 50.
- DeepWalk: DeepWalk uses truncated random walks to obtain network structure information and employs word2vec to learn node representations. In this experiment, the window size w is 5, and walk length t is 40. Hierarchical softmax is used as the optimization function.
- node2vec: we set the network is the same in [14].

The node embedding dimension of the above network embedding methods are set to 128.

2)Visualization

Fig.3 provides visualization results. In the study of embedding representation of patients, we select 1340 patients' related medical entities, i.e., diagnoses, medicines, procedures. We randomly choose patients with nine diseases, namely, Liver Diseases, Heart Failure, Atherosclerosis, Intestinal Diseases, Kidney Failure, Septicemia, Pneumonia, Gastritis and Respiratory Failure. Then, we use t-SNE to cluster patients. It can be seen from the results that the separation of these nine diseases is better.

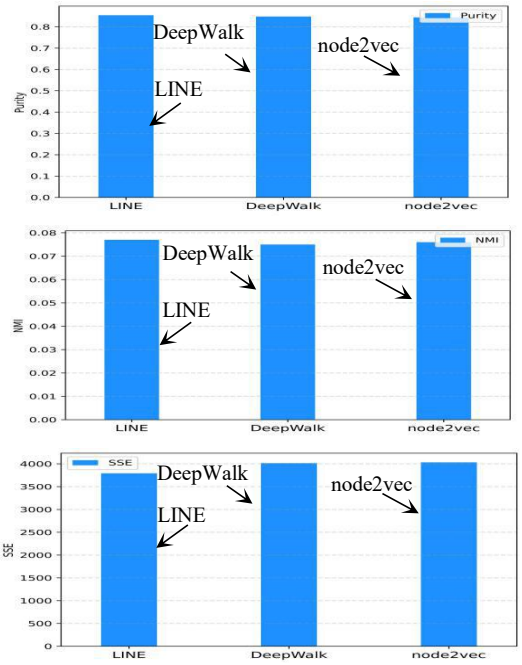
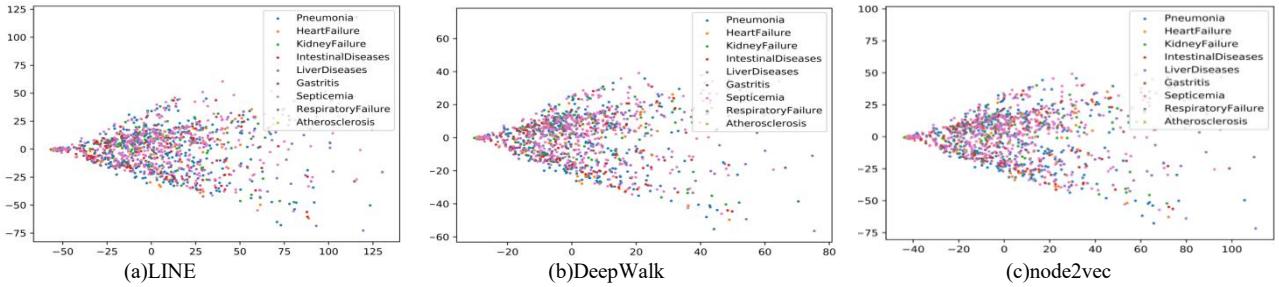


Fig. 4. Performance of different network embedding methods

We evaluate the clustering performance of the three embedding representations. We use 1340 patients and their associated medical entities as experimental data. As you can see from Fig.4, LINE outperforms the other two methods. The LINE achieves *SSE* of 3792, comparing with the node2vec and the DeepWalk is 4015 and 4032, respectively. The *Purity* and *NMI* are 0.854 and 0.077, respectively. node2vec and DeepWalk achieve 0.848, 0.844 and 0.075, 0.076, respectively.

3)Top-K Most Similar Patients

We run three network embedding methods to obtain patient representations. We randomly selected the total nine



(a)LINE (b)DeepWalk (c)node2vec

Fig. 3. Visualization of patients. Each dot indicates one patient. Color of a dot indicates the disease of the patient.

patients from nine diseases (choose one patient for each disease) to test. Table II, III and IV show top-*k* (*k*=3) similar patients in each patient nearest neighbor. From these tables, the results of the LINE and DeepWalk to find similar patients of the top-*k* of the nine test patients are consistent. However, except for three diseases (Heart Failure, Intestinal Diseases and Atherosclerosis), node2vec finds the same nearest neighbor as LINE and DeepWalk. For Heart Failure, the patient ID 782 is the second nearest neighbor in node2vec, but in the other two methods patient ID 782 is the nearest neighbor. For Atherosclerosis, the patient ID 595 in node2vec is the nearest neighbor, but in the other methods patient ID 595 is the second neighbor. For Intestinal Disease, it is not consistent with the patients obtained by the other two methods. This result might be due to the fact that they both utilize network representation learning method to learn the medical entity representations.

TABLE II
NEAREST NEIGHBOR FOR PATIENT EMBEDDING (LINE)

Patient (ID)	Nearest Neighbor (ID)	2nd Nearest Neighbor (ID)	3rd Nearest Neighbor (ID)	Disease (Label)
105	120	1167	389	Pneumonia
125	782	348	1083	Heart Failure
243	291	1255	825	Kidney Failure
389	289	36	1097	Intestinal Diseases
480	686	965	181	Liver Diseases
528	777	1056	158	Gastritis
605	1161	278	726	Septicemia
4913	1254	275	477	Respiratory Failure
6234	916	595	88	Atherosclerosis

TABLE III
NEAREST NEIGHBOR FOR PATIENT EMBEDDING (NODE2VEC)

Patient (ID)	Nearest Neighbor (ID)	2nd Nearest Neighbor (ID)	3rd Nearest Neighbor (ID)	Disease (Label)
105	120	669	1167	Pneumonia
125	348	782	255	Heart Failure
243	291	125	686	Kidney Failure
389	182	575	1191	Intestinal Diseases
480	686	965	181	Liver Diseases
528	765	158	777	Gastritis
605	1161	726	278	Septicemia
4913	1254	938	275	Respiratory Failure
6234	595	468	343	Atherosclerosis

TABLE IV
NEAREST NEIGHBOR FOR PATIENT EMBEDDING (DEEPWALK)

Patient (ID)	Nearest Neighbor (ID)	2nd Nearest Neighbor (ID)	3rd Nearest Neighbor (ID)	Disease (Label)
105	120	1167	389	Pneumonia
125	782	348	1083	Heart Failure
243	291	1255	825	Kidney Failure
389	289	36	1097	Intestinal Diseases
480	686	965	181	Liver Diseases
528	777	1056	158	Gastritis
605	1161	278	726	Septicemia
4913	1254	275	477	Respiratory Failure
6234	916	595	88	Atherosclerosis

4)The Performance of Patient Similarity

We use *HRR* to measure the performance of patient similarity. We randomly select 1500 patients and pick the most similar patient of each selected patient, and then evaluated the performance of our proposed framework with *HRR* value. Table V shows the value of *HRR*, as can be seen from Table V, LINE is superior to the other methods for measuring patient similarity. The *HRR* of LINE is 0.672, which is the best performance. Comparing to best performance, DeepWalk achieves the second best performance in *HRR*, which is 0.562, and node2vec achieves the lowest performance in *HRR*.

TABLE V
HOSPITAL READMISSION RATE(*HRR*)

Method	Technique	HRR
DeepWalk	Random Walk +Skip-gram	0.562
node2vec	Random Walk based on DFS and BFS	0.557
LINE	First-order and second-order Proximity+Negative sampling	0.672

LINE is obviously superior to the other two network embedding methods. Next, we use LINE to do comparative experiments.

5)Comparisons with Other Patient Similarity Methods

In the following experiments, LINE is used for network embedding. To evaluate the correctness and effectiveness of our method T-PS, we compare our method with the following baselines:

- baseline 1: This method considers the co-occurrence relationships between medicines when constructing the temporal medical entity association graph, and describes patients through medicines.
- baseline 2: This method considers the co-occurrence relationship between medicines and diagnoses when

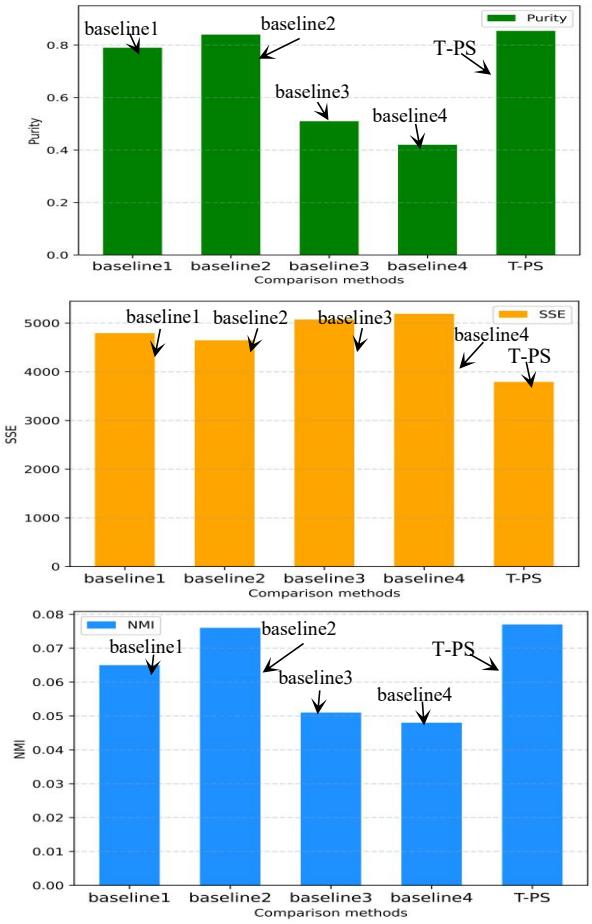


Fig.5. Performance of different methods

TABLE VI
DISEASE CLASSIFICATION RESULTS

Time Decay Function	Time Decay Function Formula	Macro-AUC	Accuracy	Macro-F1
Guassian Kernel	$\Gamma(\Delta t) = \exp\left[\frac{-\Delta t^2}{2\sigma^2}\right]$	0.514	0.792	0.457
Passage Kernel	$\Gamma(\Delta t) = \begin{cases} 1 & if\ \Delta t \le \sigma \\ 0 & otherwise \end{cases}$	0.608	0.698	0.426
Cosine Kernel	$\Gamma(\Delta t) = \begin{cases} \frac{1}{2}[1+\cos(\frac{\Delta t \cdot \pi}{\sigma})] & if\ \Delta t \le \sigma \\ 0 & otherwise \end{cases}$	0.759	0.862	0.534

constructing the temporal medical entity association graph.

- baseline 3: This method does not consider timeliness and uses medicines, diagnoses and procedures to describe patients.
- baseline 4: Code Sum based Matching(CSM) [46] obtains the patient representations by adding all medical codes vectors. Firstly, CSM uses Word2Vec to learn medical codes vectors. Then, it sums up the medical codes vectors of the patient to retrieve a single embedding. Finally, the patient similarity score is calculated by cosine similarity.

We summarize the results on the clustering task in Fig.5. The results of T-PS exhibit better performance. The CSM achieves the worst performance. A possible reason could be that the CSM method applies word embedding algorithm and ignores the latent information on the dataset. Our method represents the dataset as a graph and the learned patient representations are more accurate.

6)Impact of Time Decay Function

In order to evaluate the influence of different time decay functions on the experimental results, the Gaussian kernel and Passage kernel are selected and compared with T-PS. A disease classification task was performed to compare the effects of different function. The vectors of patients with different time decay functions are obtained. Then apply MLP classification[48] on the learned patients' vectors in order to correctly diagnose the disease suffered by the patients. In addition, we use Macro-F1, Macro-AUC and accuracy to evaluate the performance of disease classification task, and use 10-fold cross-validation[49] to evaluate the results of the remaining samples without label information, randomly selected 80% of the data for learning, and 20% of the data for MLP classification test.

Comparative results of different time decay functions for disease classification are shown in Table VI. We observe that our method achieves Macro-AUC of 0.759, the

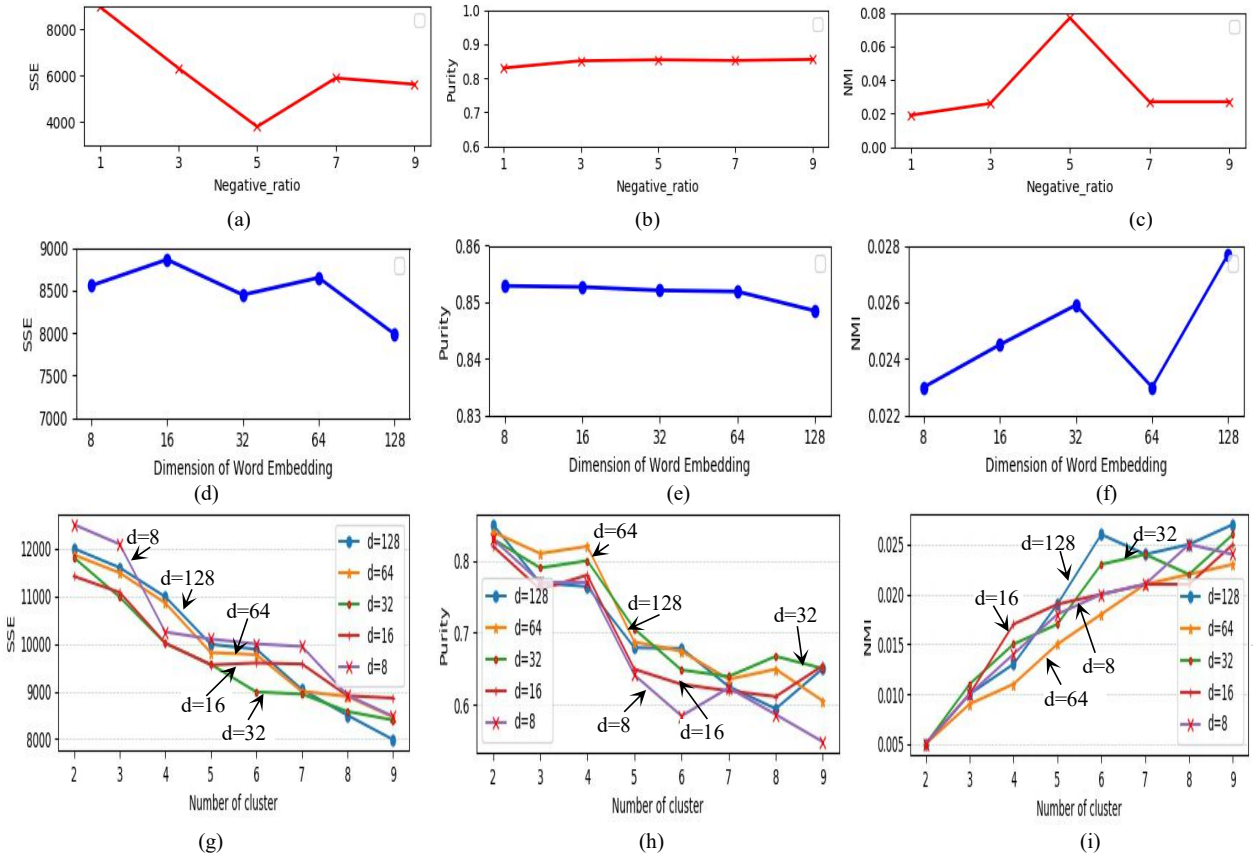


Fig 6. The performance of change embedding dimensions, the number of cluster and negative_radio. (a), (b), (c) together show the performance of changing negative_radio *r* on clustering performance. (d), (e), (f) together show the effect of embedding dimension *d* on clustering performance. (g), (h), (i) measure the efficacy of variation on the number of cluster *k*.

accuracy of 0.862, and Macro-F1 of 0.534, which outperforms all the other methods, and Gaussian kernel achieves the second highest performance. Thus, the patient representations are obtained by T-PS can enhance the performance of disease cohort classification. In general, our proposed T-PS is a good choice in practice for disease classification task due to its good performance.

7) Parameter Analysis

We study the parameter sensitivity of embedding dimension *d*, cluster number *k* and negative_ratio *r*. Fig.6 shows the performances when altering different parameters. The other parameters remain unchanged. We set *k*=9 to observe the effect of embedding dimensions on the results. *d*=32 is the smallest *SSE*. *Purity* decreases with the increase of embedding dimension *d*, and *NMI* is the largest at *d*=128. Meanwhile, when we change the number of clusters *k*, the change trend of *SSE*, *Purity* and *NMI* for different embedding dimensions *d* is consistent, *d*=128 *SSE* is the smallest, *d*=64 *Purity* is the best, and *k*=9 *NMI* is the largest. When we change negative_ratio *r*, *Purity* remains unchanged. When *r*=5, *NMI* achieves a maximum value of 0.077. The value of *SSE* decreases first and then increases, and *r*=5 is the optimal value.

VI. Conclusions and Future Work

Patient similarity is an important problem for various healthcare applications. However, due to the high-dimensional and sparse characteristics of medical data, the study of patient similarity faces many challenges. We propose a novel time-aware patient similarity framework T-PS. The framework exploits comprehensive medical information in EHRs to construct a high-quality temporal medical entity association graph. Leveraging graph-based embedding, T-PS obtain more semantic and lower dimensional patient representations to calculate patient similarity. Experimental results show that our method obtain better representations than other baselines. For future work, we will pursue to construct medical heterogeneous information network from EHRs and find more complex semantics in the network for patient similarity.

References

[1] Sun J , Wang F , Hu J , et al. Supervised patient similarity measure of heterogeneous patient records[J]. ACM SIGKDD Explorations Newsletter, 2012, 14(1):16-24.

[2] Chan, LWC, Chan, T, Cheng, LF, et al. Machine learning of patient similarity: a case study on predicting survival in cancer patient after locoregional chemotherapy[C]// Bioinformatics and Biomedicine Workshops (BIBMW), 2010 IEEE International Conference on. IEEE, 2011.

[3] Changchan Y, BuyuQ ,Jishan W, et al. Measuring patient similarities via a deep architecture with medical concept embedding[C] IEEE International Conference on Data Mining. 0.

[4] Kim M S , Clarke M A , Belden J L , et al. Usability Challenges and Barriers in EHR Training Of Primary Care Resident Physicians[J]. 2014.

[5] Srivastava, Nitish, Mansimov, Elman, Salakhutdinov, Ruslan. Unsupervised Learning of Video Representations using LSTMs[J].

[6] Choi E, Bahadori M T , Searles E , et al. Multi-layer Representation Learning for Medical Concepts[J]. 2016.

[7] JoonL ,Maslove D M , Dubin J A , et al. Personalized Mortality Prediction Driven by Electronic Medical Data and a Patient Similarity Metric[J]. PLOS ONE, 2015, 10(5): e0127428-.

[8] Peng C, Xiao W, Jian P , et al. A Survey on Network Embedding[J]. IEEE Transactions on Knowledge and Data Engineering, 2018:1-1.

[9] 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.

[10] Zhihuang Lin, and Dan Yang, "Medical Concept Embedding with Var iable Temporal Scopes for Patient Similarity," Engineering Letters, vo l. 28, no. 3, pp651-662, 2020

[11] Kasabov N, Hu Y. Integrated optimisation method for personalised modelling and case studies for medical decision support.[J]. 2010, 3(3):236-256.

[12] Tang J , Qu M , Wang M , et al. LINE: Large-scale information network embedding[J]. 24th International Conference on World Wide Web, WWW 2015, 2015.

[13] Perozzi, Bryan, Al-Rfou, Rami, Skiena, Steven. DeepWalk: Online Learning of Social Representations[J].

[14] Grover, Aditya, Leskovec, Jure. node2vec: Scalable Feature Learning for Networks[J].

[15] Daixin Wang, Peng Cui, Wenwu Zhu. Structural Deep Network Embedding[C] // the 22nd ACM SIGKDD International Conference. ACM, 2016.

[16] Dong, Yuxiao, Chawla, Nitesh V, Swami, Ananthram. metapath2vec: Scalable Representation Learning for Heterogeneous Networks[C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017.

[17] Liu P, Qiu X, Huang X. Learning context-sensitive word embeddingswith neural tensor skip-gram model[C]// 2015.

[18] Yuxin Chen, Chenguang Wang. HINE: Heterogeneous Information Network Embedding[J].

[19] Chang, S., Wei Han, Jiliang Tang, Guo-Jun Qi, C. Aggarwal and T. Huang. "Heterogeneous Network Embedding via Deep Architectures." KDD '15 (2015).

[20] Huang X , Li J , Hu X . Label Informed Attributed Network Embedding[C]// Tenth Acm International Conference on Web Search & Data Mining. ACM, 2017.

[21] Hong, Richang, He, Yuan, Wu, Le,et al. Deep Attributed Network Embedding by Preserving Structure and Attribute Information[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems:1-12.

[22] Liao L , He X , Zhang H , et al. Attributed Social Network Embedding[J]. IEEE Transactions on Knowledge & Data Engineering, 2017:1-1.

[23] Yang C , Liu Z , Zhao D , et al. Network Representation Learning with Rich Text Information[C]// International Conference on Artificial Intelligence. AAAI Press, 2015.

[24] Fei W, Lee N, Hu J, et al. Towards heterogeneous temporal clinical event pattern discovery:a convolutional approach[C]// 2012.

[25] Maida J Sewitch, Karen Leffondré, Patricia L Dobkin. Clustering patients according to health perceptions: Relationships to psychosocial characteristics and medication nonadherence[J]. Journal of Psychosomatic Research, 56(3):0-332.

[26] Ni J, Liu J , Zhang C , et al. Fine-grained Patient Similarity Measuring using Deep Metric Learning[C]the 2017 ACM. ACM, 2017.

[27] Nomura, R. (2019). Source resolvability problem with respect to a certain subclass of f-divergence. In 2019 IEEE International Symposium on Information Theory, ISIT 2019 - Proceedings (pp. 2234-2238).

[28] Xu, Kun, Feng, Yansong, Huang, Songfang,et al. Semantic Relation Classification via Convolutional Neural Networks with Simple Negative Sampling[J]. Computer Science, 2015, 71(7):941-9.

[29] Chen, Long, Yuan, Fajie, Jose, JoemonM,et al. Improving Negative Sampling for Word Representation using Self-embedding Features[J].

[30] He, Yaoyao, Xu, Qifa, Wan, Jinhong,et al. Short-term power load probability density forecasting based on quantile regression neural network and triangle kernel function[J]. Energy, 114:498-512.

[31] Coppock, C.E, Slack, S.T, Buck, G.R,et al. Effect of Recutting and Plant Maturity on Kernel Passage and Feeding Value of Corn Silage[J]. 52(10):1617-1623.

[32] Li He, Hong Zhang. Kernel K-Means Sampling for Nyström Approximation[J]. IEEE Transactions on Image Processing, 2018, PP(99):1-1.

[33] Kay, A. B. Messenger RNA expression of the cytokine gene cluster, interleukin 3 (IL-3), IL-4, IL-5, and granulocyte/macrophage colony-stimulating factor, in allergen-induced late-phase cutaneous reactions in atopic subjects[J]. Journal of Experimental Medicine, 173(3):775-778.

[34] Wenbo Zhu, Zachary T Webb, KaitianMao, et al. A Deep Learning Approach for Process Data Visualization Using t-Distributed Stochastic Neighbor Embedding[J]. Industrial & Engineering Chemistry Research, 2019, 58(22).

[35] Alistair Edward William Johnson, Tom Joseph Pollard, Lu Shen,et al. MIMIC-III, a freely accessible critical care database[J]. Scientific

Data, 2016, 3:160035.
[36] Marie S L , Cesar A , Suvarna N , et al. Disease Ontology: a backbone for disease semantic integration[J]. Nucleic Acids Research(D1):D1.
[37] Li, Ying, Zhang, Saijuan, Baugh, Reginald F,et al. Predicting surgical case durations using ill-conditioned CPT code matrix[J]. Iie Transactions, 42(2):121-135.
[38] Young Donald S, Sachais Bruce S, Jefferies Leigh C. Comparative Costs of Treating Adults and Children within Selected Diagnosis-related Groups[J]. Clinical Chemistry, 2020(1):1.
[39] Hanna J , Joseph E , Mathias Brochhausen. Building a drug ontology based on RxNorm and other sources[J]. Journal of Biomedical Semantics, 2013, 4(1):44-44.
[40] Selvida D , Zarlis M , Situmorang Z . Analysis of the effect early cluster centre points on the combination of k-means algorithms and sum of squared error on k centroid[J]. IOP Conference Series Materials ence and Engineering, 2020, 725:012089.
[41] Leggio, B., Napoli, A., Nakazato, H., & Messina, A. (2020). Bounds on mixed state entanglement. Entropy, 22(1), 62. https://doi.org/10.3390/e22010062
[42] Meil, Marina. Comparing clusterings---an information based distance[M]. Academic Press, Inc. 2007.
[43] Joon L , Maslove D M , Dubin J A , et al. Personalized Mortality Prediction Driven by Electronic Medical Data and a Patient Similarity Metric[J]. PLoS ONE, 2015, 10(5):e0127428.
[44] Le Q V , Mikolov T . Distributed Representations of Sentences and Documents[J]. 2014..
[45] PichottaK , Mooney R J . Learning Statistical Scripts with LSTM Recurrent Neural Networks[C]// Thirtieth Aaai Conference on Artificial Intelligence. AAAI Press, 2016.
[46] Choi, Edward , et al. "Medical Concept Representation Learning from Electronic Health Records and its Application on Heart Failure Prediction." (2016).
[47] Keikha M , Gerani S , Crestani F . Time-based relevance models[C]// Proceeding of International Acm Sigir Conference on Research & Development in Information Retrieval. ACM, 2011.
[48] Kim B , Lee S M , Seo J K . Improving learnability of neural networks: adding supplementary axes to disentangle data representation[J]. 2019.
[49] Zhang H , Yang S , Guo L , et al. Comparisons of isomiR patterns and classification performance using the rank-based MANOVA and 10-fold cross-validation[J]. Gene, 2015, 569(1):21-26.