World Wide Web (2020) 23:2023–2042
https://doi.org/10.1007/s11280-020-00781-3

Medical treatment migration behavior prediction and recommendation based on health insurance data

Lin Cheng[1] · Yuliang Shi[1,2] · Kun Zhang[1,2]

Received: 9 September 2019 / Revised: 23 December 2019 / Accepted: 2 January 2020 /
Published online: 12 March 2020

Abstract

How to accurately predict the future medical treatment behaviors of patients from the historical health insurance data has become an important research issue in healthcare. In this paper, an Attention-based Bidirectional Gated Recurrent Unit (AB-GRU) medical treatment migration prediction model is proposed to predict which hospital patients will go to in the future. The model considers the impact of medical visit on the future medical behavior, on the basis of Bidirectional Gated Recurrent Unit (B-GRU) framework, we introduce an attention mechanism to determine the strength of hidden state at different moments, which can improve the predictive performance of the model. Due to medical treatment in different places has an important impact on the distribution of health insurance funds, the individual patient would be expected to the appropriate hospital and get the appropriate medical treatment. Therefore, when medical treatment prediction has been completed, this paper proposes a Similarity and Double-layer CNN-based (SD_CNN) medical treatment migration recommendation model. The model introduces a CNN framework to achieve patient similarity learning, and compares similarities to recommend whether patients need medical treatment migration. Finally, the experiment demonstrates that the model proposed in this paper is more accurate than other models.

Keywords health insurance data · medical treatment migration prediction · medical treatment recommendation

Guest Editors: Jie Shao, Man Lung Yiu, and Toyoda Masashi

✉ Yuliang Shi
shiyuliang@sdu.edu.cn

Lin Cheng
chenglin123_sdu@163.com

Kun Zhang
kunzhangcs@126.com

Extended author information available on the last page of the article

1 Introduction

Medical treatment migration generally can be simply defined as the behavior of seeking medical treatment outside the insured areas by insured persons. In the process of medical treatment, a large number of data will be generated, such as the choice of medical hospitals, the number of medical treatments, and medical expenses information. These data constitute health insurance data. Through the health insurance data, we can fully understand the patients' medical treatment. It includes medical visit sequences over time, where each medical visit includes a medical code, diagnosis, medical hospital, medical items, expenses, and so on. It has a wide range of applications in the field of health insurance research, such as risk prediction [6, 14, 16, 22], disease prediction [5, 17, 27]. In recent years, with the rapid expansion of the number of floating populations in China, the number of people who migrates to different hospitals to seek medical treatment has also increased. Due to the differences in medical levels between hospitals, more patients are willing to go to hospitals with higher medical level, resulting in the waste of hospital resources and unreasonable distribution of health insurance funds. How to accurately predict the future medical treatment behavior of patients has become an important research issue in healthcare of China.

A traditional method of prediction in the field of health insurance is to treat each patient's visit as a feature vector and to use that vector as an input to construct a predictive function. Given the features, training data are used to fit the predictive function to minimize an appropriate cost function. Currently, in order to construct predictive models using medical visit sequences, Recurrent Neural Networks (RNNs) are widely used [10, 11]. However, RNNs cannot effectively address long-term dependencies. When the patient's visit sequence is too large, the predictive performance of the RNNs model will decrease. Moreover, RNNs also ignore the effects of time intervals on medical treatment behavior.

At present, the recommendation methods in healthcare mainly take two steps: (1) Calculate the similarity between patients. (2) Generate a recommendation list for the patient based on the similarity and patient history of medical treatment. Based on health insurance data, many deep learning methods have been widely adopted and rapidly developed in patient similar learning [15, 28, 32, 33], such as automatic encoders, Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNN). In [40], CNN has proven its superior ability to measure patient similarity. However, one drawback of the traditional CNN framework is that it does not make full use of time and context information in health insurance data for similarity learning. Therefore, how to use the health insurance data with high-dimensional time characteristics for similarity learning has great challenges.

In response to the above issues and challenges, in this paper, we aim to solve the following key problems: how to consider the impact of time information on medical behavior to improve prediction performance; how to recommend whether the patient has a medical migration. To tackle these problems, in the process of medical migration prediction, an Attention-based Bidirectional Gated Recurrent Unit (AB-GRU) medical treatment migration prediction model is proposed. Under the condition that the medical migration prediction has been completed, this paper proposes a CNN-based medical treatment migration recommendation model. On the basis of CNN framework, a matching matrix is introduced to achieve similarity learning among patients, and compares the similarity to recommend whether patients need medical treatment migration. In summary, our contributions are as follows:

- Considering the impact of time information within the visit sequences, we propose a medical treatment migration prediction model in a bidirectional GRU framework to predict which hospital patients will go to in the future.
- In order to quantify the impact of each medical visit on the future hospitalization behavior, on the basis of bidirectional GRU framework, we introduce an attention mechanism. The attention score is used to determine the strength of the hidden stateat different moments, which can improve the prediction performance of the model.
- In order to realize medical treatment migration recommendation, on the basis of CNN framework, this paper introduces a matching matrix for similarity learning, and realizes the medical migration recommendation through the similarity comparison, thatis, whether the patient needs medical treatment migration.

The rest of this paper is organized as follows: In section 2, we introduce relevant works on medical treatment behavior and similarity learning. Section 3 presents the details of the medical treatment migration prediction and recommendation model. The experimental results are presented in section 4. Section 5 is a summary of the work in this paper.

2 Related work

In this section, we mainly introduce the research related to medical treatment behavior and similarity learning.

2.1 Medical treatment behavior

As far as the current research on medical treatment behaviors of is concerned [35], it mainly focuses on two aspects. Firstly, the research on the types and characteristics of medical treatment behaviors [31, 37], mainly concentrated on method and the choice of hospitalization behavior. Secondly, the research on the influencing factors of medical treatment behavior [1, 4], mainly from the aspects of personal attributes, economic factors, social factors, etc., and obtained rich research results.

Bei et al. used a statistical survey to study the behavior of medical treatment, and predicted patients' medical behavior by Multivariable Logistic Regression algorithm [3]. Wang et al. provided a computational framework for studying health disparities among cohorts based on individual level features, such as age, gender, income, etc. This framework to find health disparities among subpopulations in an influenza epidemic and evaluate vaccination prioritization strategies to achieve specific objectives [29]. Lu et al. studied the influencing factors of the choice of hospitalization behaviors among agricultural transfer [20]. That is, they analyzed the influencing factors of the choice behavior of medical treatment, and predicted the choice of hospitalization behaviors after illness by regression algorithms. In addition, Duan et al. proposed combining the gray correlation analysis method with the multi-class selection model to realize the research and prediction of the community residents' medical treatment behavior [13]. Zheng et al. used social action theory and an analytical model to examine the main influencing factors of rural residents' medical treatment, and constructed a model to study the behavior of rural residents [39]. However, these research methods ignored the time information inside the medical treatment sequence.

Currently, using medical visit sequences, recurrent neural networks are widely used. Baytas et al. proposed a novel Time-Aware LSTM to handle irregular time intervals in longitudinal visit records [2], Che et al. developed novel deep learning models for multivariate time series with missing values [7]. But, these methods did not achieve the interpretability of models.

2.2 Similarity learning

Suo et al. proposed a deep similarity learning framework based on the CNN, and simultaneously implements patient feature learning and similarity metrics [27]. In [36], for the high-dimensional, heterogeneous and complex characteristics of medical insurance data, Zhan et al. proposed a new similarity learning method, namely the generalized Mahalanobis similarity function with pairwise constraints. At the same time, considering that there are always some non-discriminatory features and contain redundant information, the author encoded the low rank structure as the similarity function to perform feature selection. To address the data sparsity issue, He et al. developed a novel similarity metric to measure the similarity between two set of trajectories so as to validate whether the reconstructed trajectory set can well represent the original traces [15]. In [25], in order to realize the construction of the standard bibliographic topic model, a standard bibliographic recommendation method based on the fusion multi-feature theme model was proposed by Shao et al. However, these similarity learning methods not only ignored the influence of time attributes and context information in test data on similarity learning, but also did not consider the data imbalance problem.

Compared with the above mentioned methods, the model proposed in this paper not only considers the impact of historical medical visit sequences on future medical treatment behavior, but also considers the influence of temporal attributes and context information on similarity learning.

3 Methods

3.1 Basic symbols

We denote all the unique medical visit codes from the health insurance data as c_1, c_2,..., $c_{|C|} \in C$, where $|C|$ is the number of unique medical visit codes. Assuming there are N patients, the n^{th} patient has $T^{(n)}$ visit records in the health insurance data. The patient can be represented by a sequence of visits X_1, X_2,…,$X_{T^{(n)}}$. Each visit X_i contains a set of feature vectors $x \in \mathbb{R}^{|C|}$. Therefore, each patient's visit can be viewed as a matrix, where the horizontal dimension corresponds to medical events and the vertical dimension corresponds to visits. The $(i,j)^{th}$ entry of a matrix is 1 if code c_j is observed at time stamp V_i for the corresponding patient. Since the number of visits of different patients varies, we pad zero to the visit dimension, making each patient have a fixed length of visits $\mathrm{t} = \max\{V_i\}_{i=1}^{T_n}$, for the sake of CNN operations.

3.2 Overall process

The main goal of this paper is to predict which hospitals will be moved to the hospital in the future and whether medical treatment migration is needed.

Figure 1 depicts the entire process of the model. In terms of medical migration prediction, the model firstly uses the Grey Relational Analysis (GRA) method to analyze the influencing factors. In order to quantify the impact of each medical visit sequence on future medical treatment behavior, we propose AB-GRU. AB-GRU employs bidirectional gated recurrent unit to remember all the information of both the past visits and the future visits, and it introduces an attention mechanism to measure the relationships of different visits for prediction. When medical treatment prediction has been completed, in order to judge whether the patient is worthy of medical treatment migration, on the basis of CNN, this paper constructs a similarity learning framework to respectively calculate patient similarity from the migrated and non-migrated patients, and obtain Group A and Group B according to the similarity. Then, we can obtain the output result y by the Multiple Logistic Regression (MLR) function. By comparing y, it is recommended whether the patient is worthy of medical treatment migration.

3.3 Medical treatment migration prediction

3.3.1 Feature selection

In healthcare, there are many factors of affecting medical treatment migration as shown in Table 1. A great number of features are collected before building the predictive model but not all the variables are informative and useful. It is imperative to eliminate the redundancy of the features and select more informative variables for increasing the accuracy and efficiency of the predictive model. In our model, we use the Grey Relational Analysis (GRA) method to analyze the influencing factors. GRA aims to determine whether they can better distinguish target instances than other features by calculating the degree of association [13, 38]. Then, the process of feature selection is as follows:

1) Normalize the original health insurance data X by the min-max standardization method, and the calculation is as follows:

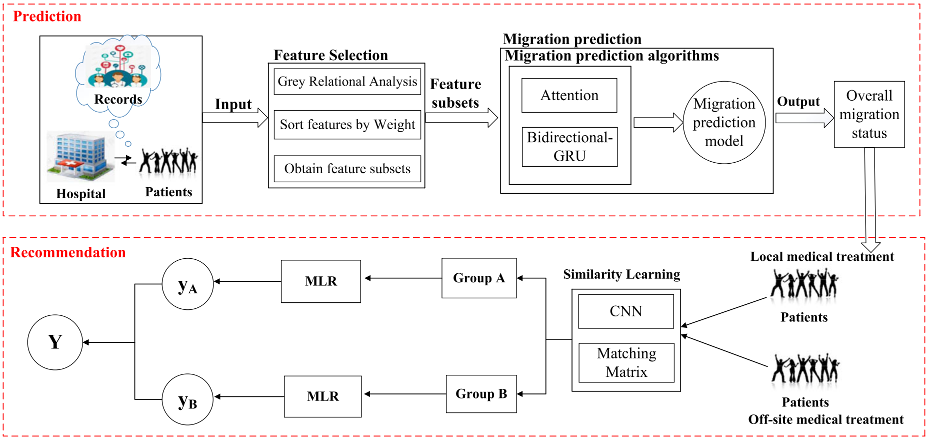

Fig. 1 The overall process

Table 1 Factors about medical treatment in our dataset

Categories	Factors	Description
Patients	Age	The insured person's age
	Gender	Male = 1, Female = 2
	Income	The patient's income
	Insured Category	Staff = A, Resident = B
	Insured place	Insured place of patients
	Disease category	21 categories
	Distance	Distance between the insured and the hospital
	Industry category	Insured person's industry (9 categories)
Hospital	Hospital Name	Name of medical hospital
	Hospital level	Hospital level: 1/2/3
	Average Hospitalization Days	The complexity of hospital treatment of certain diseases
	Average Cost	The complexity of hospital treatment of certain diseases
	Maximum number	The maximum number of hospital in a period of time

$$x^{'} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

2) Suppose that the system reference sequence after the mean value change of each feature in the original sample is $X_0 = (x_0(1), x_0(2), \cdots, x_0(n))$, whether medical treatment is a comparative sequence $X_k = (x_k(1), x_k(2), \cdots, x_k(n))$, k = 1, 2, $\cdots$, m. Then, the correlation coefficient between X_0 and X_k is calculated by eq. (2):

$$\delta_i(k) = \frac{min_i min_k |x_0(k) - x_i(k)| + \rho max_i max_k |x_0(k) - x_i(k)|}{|x_0(k) - x_i(k)| + \rho max_i max_k |x_0(k) - x_i(k)|} \tag{2}$$

where $\delta_i(k)$ represents the correlation coefficient of x_i to x_0 in the k^{th} data. ρ represents the resolution coefficient (ρ=0.5).

3) The relationship θ between each factor X_0 and whether or not to migrate X_k:

$$\theta_i = \frac{1}{n} \sum_{k=1}^{n} \delta_i(k) \tag{3}$$

Finally, we can sort the relevance degree θ and select the main feature set that affects the medical migration according to the sorting size, which lays a foundation for the construction of the medical migration prediction model.

3.3.2 Migration prediction

In this section, we propose an Attention-based Bidirectional GRU medical treatment migration prediction model (AB-GRU) to achieve the migration prediction of the patients.

The goal of the proposed model is to predict the $(t+1)^{th}$ visit's hospital. As shown in Fig. 2, given the visit information from time 1 to t, the i^{th} visit x_i can be embedding into a vector v_i. The vector v_i is fed into the Bidirectional GRU, which outputs a hidden state H_i. Along with the set of hidden state H_i, we can compute the degree of correlation a_i between the hidden state and medical treatment behavior at each moment by attention operation. Finally, from the visit sequence v_i and hidden state H_i at all time, we can get the final prediction through the softmax function.

Embedding layer Given a visit sequences $X_i(\mathrm{i}=1,2,\ldots,|C|)$. We can get its vector representation $v \in \mathbb{R}^K$, as follows:

$$v = A^T x \tag{4}$$

where K represents the dimension of the embedding layer, $A \in \mathbb{R}^{|C| \cdot K}$ is the weight matrix.

Bidirectional GRU As a variant of the standard recurrent neural network (RNN), the gated recurrent unit (GRU) was originally proposed [18]. For each position t, GRU computes h_t with input x_t and previous state h_{t-1}, as:

$$r_t = \beta(W_r x_t + U_r h_{t-1}) \tag{5}$$

$$\pi_t = \beta(W_\pi x_t + U_\pi h_{t-1}) \tag{6}$$

$$\widetilde{h}_t = tanh(W_c x_t + U(r_t \odot h_{t-1})) \tag{7}$$

$$h_t = (1 - \pi_t) \odot h_{t-1} + \pi_t \odot \widetilde{h}_t \tag{8}$$

where h_t, r_t and π_t are d-dimensional hidden state, reset gate, and update gate, respectively. W_r, W_π, W_c and U_r, U_π, U are the parameters of the GRU. β is the sigmoid function, and $\odot$ denotes element-wise production.

A bidirectional GRU consists of a forward and backward GRU. The forward GRU $\overrightarrow{h}_T$ reads the input visit sequence from x_1 to x_T and calculates a sequence of forward hidden states $\left(\overrightarrow{h}_1, \overrightarrow{h}_2, \cdots, \overrightarrow{h}_T\right)$. The backward GRU $\overleftarrow{h}_T$ reads the visit sequence in the reverse order from x_T to x_1, resulting in a sequence of backward hidden states $\left(\overleftarrow{h}_1, \overleftarrow{h}_2, \cdots, \overleftarrow{h}_T\right)$. We can obtain the final hidden state H_t, as follows:

$$H_t = \left[\overrightarrow{h}_T, \overleftarrow{h}_T\right] \tag{9}$$

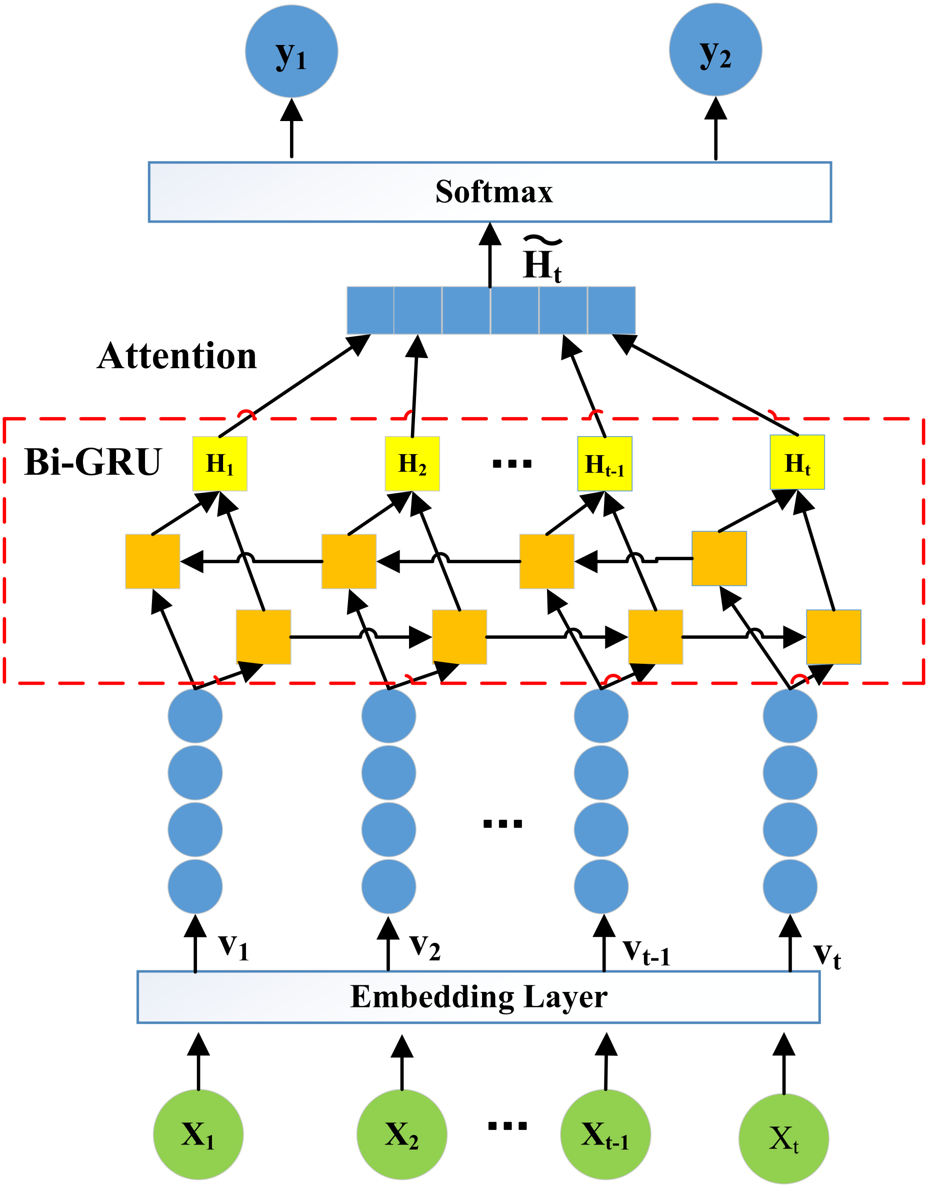

Fig. 2 The process of AB-GRU

Attention mechanism The core idea of the attention mechanism is to assign more attention to important content and less attention to other parts [8, 12, 19, 21, 23]. In the process of medical treatment migration prediction, the traditional neural networkmodel ignores the impact of the length of the time interval within the visit sequences on the modeling, since the contribution of each visit to the current moment is not necessarily the same. Therefore, considering that not all features contribute to the prediction, we add the attention layer to the bidirectional GRU framework. The attention score is used to determine the strength of the hidden state during the modeling process of the medical treatment sequence, thereby significantly improving the modeling ability of the prediction model.

Implemented in the Attention mechanism as follows:

$$u_t = tanh(WH_t + b) \tag{10}$$

$$a_t = \frac{exp\left(u_t^T u\right)}{\sum_{t=1}^{T} exp\left(u_t^T u\right)} \tag{11}$$

$$\gamma = \sum_{t=1}^{T} a_t H_t \tag{12}$$

where $W \in \mathbb{R}^{L \times |C|}$ and $b \in \mathbb{R}^L$ are corresponding weights and bias vectors, u_t is the importance vector. a_t represents the normalized weight by eq. (11). γ is the weighted sum of each H_t with a_t.

The output of the attention layer $\widetilde{H}$ is:

$$\widetilde{H} = \sum_{n=1}^{N} \gamma \tag{13}$$

Finally, $\widetilde{H}$ is fed through the softmax layer to produce the $(t+1)^{th}$ choice of the medical treatment behaviors defined as:

$$y = softmax\left(W_c\widetilde{H} + b_c\right) \tag{14}$$

where $W_c \in \mathbb{R}^{2L}$ and $b_c \in \mathbb{R}^L$ are the parameters to be learned.

Interpretation In healthcare, we need to understand the clinical meaning of each dimension of visits, and analyze which visit are crucial to the medical treatment migration prediction.

In our proposed AB-GRU model, the attention can be used to assign weights to the hidden state of each visit. It is easy to find the importance of each medical visit by analyzing the weight of each medical visit. We sort the weight of each dimension in the hidden state in reverse order, and then select the top K weights, as shown below:

$$argsort(a_t[:, n])[1 : K] \tag{15}$$

where $a_t[:, n]$ represents the attention weight of each dimension in the t^{th} visit. By analyzing the top K visits, we can obtain which visits have an important impact on the migration prediction. Detailed examples and analysis are given in Section 4.2.4

3.4 Medical treatment migration recommendation

In this section, under the premise that the medical prediction has been completed (completed in Section 3.3), we propose a CNN-based medical treatment migration recommendation model to judge whether patients need medical treatment migration.

3.4.1 Similarity learning

In this section, we mainly use the CNN framework to achieve patients similarity learning [27]. Figure 3 depicts a CNN-based patient similarity learning framework. The framework firstly maps the one-hot feature matrix of patient A through the Embedding layer to a low-dimensional sparse matrix. Convolution and maximum pooling are applied to each matrix. The eigenvectors of the matrices are then aggregated to form a composite vector. Patient B shares the same embedding and CNN parameters. The composite vector of patient A and patient B obtains the similarity feature vector through the matching matrix and the conversion layer. Finally, the similarity feature vector is used to obtain the similarity probability of patient A and patient B through the softmax layer.

Embedding layer The original health insurance data is high-dimensional and sparse. To reduce feature dimensions and learning relationships among medical sequences, we use a ReLU function to embed feature matrices into a vector space. Each medical visit x_i is mapped to a vector $v_i \in \mathbb{R}^d$ with the following equation:

$$v_i = ReLU(W_v x_i + b_v) \tag{16}$$

where d represents the embedding dimension, W_v and b_v are weight matrix and bias vector to be learned. After the embedding operation, we were able to obtain the embedding matrix $V \in \mathbb{R}^{t \times d}$ for each patient.

Convolution layer The convolutional layer has p different filter sizes and the number of filters per size is q, so that the total number of filters is m = pq. Each filter is defined as $W_e \in \mathbb{R}^{h \times d}$, where h is a window size of visit length, meaning that the convolution operation is applied over h sequential timestamps. Suppose a filter is applied over a concatenation from visit vector v_i to v_{i+h-1}, a feature representation D_i is generated using $D_i = ReLU(W_e \cdot v_{i+h-1} + b_e)$. This filter is applied to each window of timestamps $\{v_{1:h}, v_{2:h+1}, \cdots, v_{t-h+1:t}\}$ with a stride equal to 1, to produce a feature map $D = \{D_1, D_2, \cdots, D_{t-h+1}\}$, where $D \in \mathbb{R}^{t-h+1}$. Since we have totally m filters, we can obtain m feature maps. The outputs from the convolutional layer are then passed into the pooling layer. A max pooling is applied over c as $\widetilde{D} = max\{D\}$, where $\widetilde{D}$ is the maximum value corresponding to a particular filter. The key idea here is to capture the most important feature for each feature map. It can naturally deal with variable visit lengths,

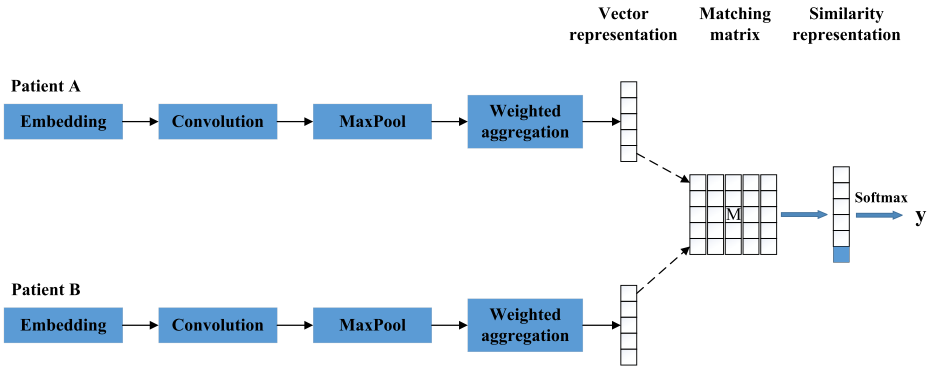

Fig. 3 Patient similarity learning framework based on CNN

since the padded visits have no contribution to the pooled outputs. The pooled outputs from all the filters are concatenated to form a vector representation $z \in \mathbb{R}^m$. z is the vector representation of the original embedding matrix V.

Similarity learning In order to realize the similarity learning of patient A and patient B, a matching matrix M is introduced. The matching matrix $M \in \mathbb{R}^{m \times m}$ is a symmetric matrix with m rows and m columns, which is used to convert the similarity vector of patient A and patient B. So, the similarity between patient A and patient B can be measured by eq. (17):

$$S = z_A M z_B \tag{17}$$

where z_A represents the vector representation of patient A; z_B represents the vector representation of patient B.

To ensure the symmetric constraint of M, it is decomposed as $M = L^T L$, where $L \in \mathbb{R}^{g \times m}$, with $g < m$ to ensure a low rank characteristic. we consider the symmetric constraint and convert patient vectors to get a similarity vector, as to ensure that the order of patients has no effect on the similarity score. We first convert z_A and z_B into a single vector with their dimension holds using the eq. (18):

$$Z = W_h z_A \oplus W_h z_B \tag{18}$$

where $W_h \in \mathbb{R}^{m \times m}$ and $\oplus$ is a bitwise addition.

After that, Z and S are concatenated and then fed into a fully connected softmax layer, to get an output probability y', indicating the similarity degree between two patients.

$$y' = softmax\left(W_f[Z; S] + b_f\right) \tag{19}$$

3.4.2 Medical migration recommendation

In order to judge whether the patient needs medical treatment migration, that is, to seek medical treatment outside the insured areas. We calculate patient similarity from the migrated and non-migrated patients according to the method in Section 3.4.1, and find the top K patients with the highest similarity from the migrated and non-migrated people respectively to obtain Similarity Group A and Similarity Group B. According to the Historical medical treatment behaviors of Group A and Group B, we respectively obtain the vector representations W_{id} of Group A and Group B through the embedding layer and convolution layer operations described in Section 3.4.1, and then we obtain the output y_i through the Multiple Logistic Regression (MLR) function, which represents the ability of Group A and Group B to treat certain diseases in their hospitals.

$$y_i = MLR(W_{id}x) + b_d \qquad i = A, B \tag{20}$$

$$Y = max\{y_i\} \qquad i = A, B \tag{21}$$

By comparing y_i, it is recommended whether the patient is worthy of medical treatment migration.

4 Experiments

In this section, we evaluate our proposed model on real dataset, which is more accurate than other methods. Note that in order to protect the privacy and safety of patients and hospitals, we anonymize the corresponding experimental data. Among them, we use Hospital-A, Hospital-B and other forms to represent the name of the hospital.

4.1 Data description

In this section, we evaluate our proposed model on real data. The dataset comes from a certain area of China from 2012 to 2017. We picked out 498,080 medical records, which contain 31,130 patients who have migrated, 44 hospitals, and 21 diseases (classified according to ICD-10 standard). Table 2 describes the statistics about the dataset.

4.2 Medical treatment migration prediction

4.2.1 Experimental setup

To evaluate the accuracy of predicting the hospital for the next visit, we used a measure for prediction task: (1) Accuracy, which is the ratio of the predicted results equal to the actual results. (2) Weighted F1-score, which calculates F1-score for eachclass and reports their weighted mean.

In this paper, we implemented all the prediction methods with the Python language. We used Adadella [34] optimizer the batch size of 500 patients. We randomly divided the dataset into training, validation and test set in a 0.75, 0.1, 0.15 ratio. We set the dimension of embedding m as 100, and the dimensionality of hidden state of GRU as 100. We used 100 iterations for each method and report the best performance.

4.2.2 Feature selection

The consequences of discriminative feature selection according Grey Relational Analysis method are presented below. Table 3 displays the weight value of all features calculated by GRA algorithm, and are ordered by weight from high to low. We setup the threshold (=0.5) to exclude some uninformative feature. In our experiments, the top 11 features obtained by GRA are considered as informative features. In addition, we can also conclude that the complexity of disease treatment in hospitals (Average Hospitalization Days and Average Cost) are more likely to affect the patient's medical treatment migration.

Table 2 Statistics of medical treatment datasets

Category	Migration situation	Total	Density
Patients	31,130	520,954	5.98%
Hospitals	44	204	21.6%
The number of disease	1295	6497	19.9%
Medical records	498,080	7,718,863	6.45%

Table 3 Factors and correlation

Factors	Correlation
Average Cost	0.9498
Average Hospitalization Days	0.9084
Hospital Level	0.8877
Income	0.8273
Age	0.7909
Treatment Rate	0.6599
Disease Category	0.6133
Gender	0.5945
Insured Category	0.5510
Maximum number	0.5133
Distance	0.5087
Industry Category	0.4766
Insured Place	0.4587

4.2.3 Migration prediction

In this subsection, we compare it to several prediction methods in order to evaluate the predictive performance of our proposed model in migration prediction. The methods are described as follows:

MLR: This is the traditional Multiple Logistic Regression model.

Navie Bayes: This is a Classification Method Based on Bayes' Theorem and Characteristic Condition Independent Hypothesis.

SVM: This is a supervised learning algorithm used to solve classification problems.

RNN: This is the traditional unidirectional Recurrent Neural Network.

BGRU: This model uses only bidirectional GRU to predict future medical information without using any attention mechanisms.

AB-GRU (our prior work) [9]: An attention-based bidirectional GRU prediction model.

Table 4 shows the accuracy and weighted F1-score of methods in migration prediction. It can be concluded from Table 4 that the prediction performance of AB-GRU proposed in this paper is better than other prediction methods. It is because that the AB-GRU considers the impact of historical medical visits and time information on future medical migration.

Figure 4 describes the ROC of several models. From Fig.4, we can conclude that the area of ROC of AB-GRU proposed in this paper is the largest, that is, the AUC value is the largest. So AB-GRU in this paper has better predictive performance.

Table 4 The accuracy and weighted F1-score of methods in migration prediction

Methods	Accuracy	F1-score
MLR	0.6855	0.6118
Naive Bayes	0.6783	0.6074
SVM	0.7096	0.6285
RNNs	0.7445	0.7049
BGRU	0.7678	0.7194
AB-GRU	0.8068	0.7423

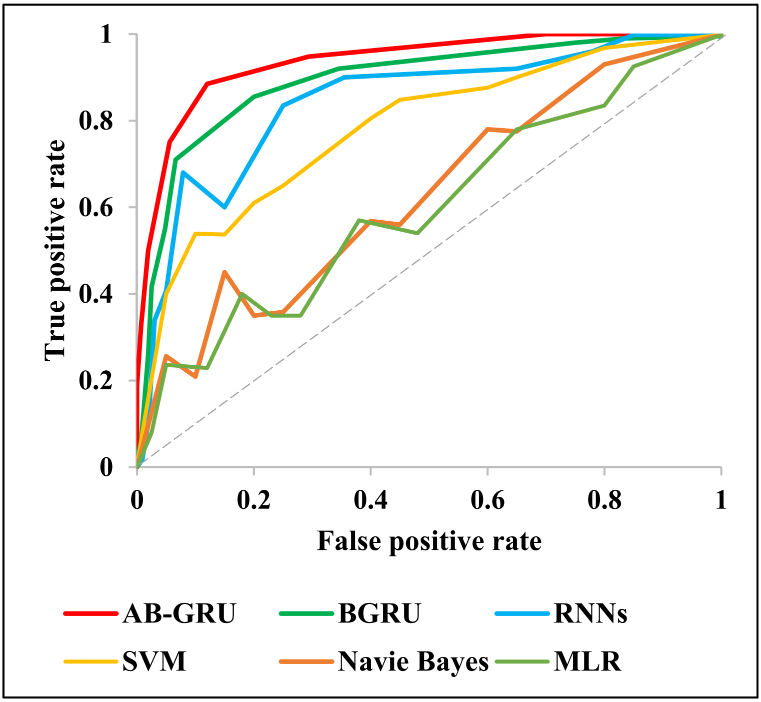

Fig. 4 The compassion of ROC

4.2.4 Model interpretation

In this subsection, we discuss interpretability of the predictive results. For each visit we are able to calculate the attention weight associated with it using eq. (11). We first select two patients with tumor disease, and we illustrate how our model utilizes the information in the patient visits for prediction. For each patient, we show each visit, the time stamp of each visit, medical hospital during each visit, and the weight assigned by AB-GRU to each visit.

From Table 5, we can observe that the weight assigned to most of the visits of patient 1 is close to 0, such as visit 1, visit 4 and visit 5, which means that they are ignored during prediction. However, for visit 2 and visit 3, although the two visits occurred in long days ago, they are assigned a larger weight, indicating that they have a long-term impact on the prediction of future medical institutions. Therefore, the predictive result is Hospital-B, which is mainly affected by visit 2 and visit 3.

Table 5 Patient 1-visit records

Medical visit records	Average weight of hidden state	Target hospital
Visit1 (305 days ago)	0.0068	Hospital-A
Visit2 (195 days ago)	0.2348	Hospital-B
Visit3 (111 days ago)	0.7509	Hospital-B
Visit4 (41 days ago)	0.0018	Hospital-A
Visit5 (14 days ago)	0.0057	Hospital-A
Prediction	Hospital-B (Actual)	Hospital-B(0.8034)

From Table 6, we can see that for patient 2, this weight is very useful for interpretation since our model focuses on visits with nonzero weights. For visit 4–5, which occurred within the last two months, AB-GRU gives large weights to the hidden of state of last two visits, indicating that the last two visits have a larger impact on choice of the hospitalization behavior of patient 2. So, the predictive result is Hospital-B, which is mainly affected by visit 4 and visit 5.

4.3 Medical migration recommendation

4.3.1 Experimental setup

To verify the performance of the patient similarity approach presented in this paper, we will compare the following methods.

Euclidean and Cosine: The similarity between samples is measured by calculating the Euclidean and cosine distances. These two methods directly measure the similarity of the original data space without learning any mapping parameters.

LMNN [30] is a classic metric learning method that brings the k nearest neighbors of the same class closer together and can separate the examples of different categories on a large scale.

GMML [24] represents the learning process as an unconstrained smooth and convex optimization problem.

K-means is a traditional clustering algorithm.

We evaluated the results of similarity learning by using three widely used criteria, the Rand Index (RI), Purity and Normalized Mutual Information (NMI) [26].

RI represents the percentage of the correct decision, calculated as follows:

$$RI = \frac{a+b}{\binom{n}{2}} \tag{22}$$

where a is the number of pairs belonging to the same group, b is the number of pairs from different groups, and n is the total number of patients. Generally, the higher the RI, the better the similarity learning effect.

The equation for calculating purity is as follows:

$$Purity(Cluster, Cohort) = \frac{1}{n}\Sigma_i max_j \left| p_i \cap q_j \right| \tag{23}$$

where $Cluster = \{p_1, p_2, \cdots, p_i\}$ is a collection of clusters, $Cohort = \{q_1, q_2, \cdots, q_j\}$ is a set of

Table 6 Patient 2-visit records

Medical visit records	Average weight of hidden state	Target hospital
Visit1 (175 days ago)	0.0016	Hospital-B
Visit2 (124 days ago)	0.0047	Hospital-B
Visit3 (74 days ago)	0.0277	Hospital-B
Visit4 (50 days ago)	0.3932	Hospital-B
Visit5 (28 days ago)	0.5728	Hospital-B
Prediction	Hospital-B (Actual)	Hospital-B(0.8849)

classes or queues. The upper limit of purity is 1, indicating a perfect match between the partitions.

NMI measures the shared information of two groups, and the equation is as follows:

$$NMI(Cluster, Cohort) = \frac{I(Cluster, Cohort)}{[Q(Cluster) + Q(Cohort)]/2} \tag{24}$$

where I represents the mutual information of two random variables and Q is the information entropy of a given random variable. The value of NMI varies between 0 and 1. When the similarity grouping is the same, I reaches the maximum value of 1.

We firstly train the similarity model described in Section 3.4 to obtain optimized CNN parameters and matching matrices. Then use the similarity framework to calculate and rank the similarity of the training data. Finally, we validate the model by performing a medical behavior recommendation. The data set was randomly divided into training set, validation set, and test set (0.75:0.1:0.15) for similarity training.

In the training process, the similarity learning framework is implemented using TensorFlow. Adadella [34] is used to optimize model parameters. Unlike the normal CNN model entered as a small batch of patients, the similarity framework is trained on a batch of patient pairs to ensure that each patient pair can be measured.

4.3.2 Similarity comparison

In Table 7, we use the RI, Purity and NMI to measure the performance of the similarity learning algorithm. A higher value means more consistency between the grouping and the real label, more similar samples are grouped together, indicating better similarity learning performance. From Table 7 we can conclude that the performance of the method proposed in this paper is significantly better than other methods. We denote the proposed framework in Section 3.4 as CNN_triple.

4.3.3 Performance comparison

In this subsection, we compare it to several methods in order to evaluate the performance of our proposed model. The methods are described as follows:

MLR: This is the traditional Multiple Logistic Regression model.

Navie Bayes: This is a Classification Method Based on Bayes' Theorem and Characteristic Condition Independent Hypothesis.

SVM: This is a supervised learning algorithm used to solve classification problems.

Table 7 Comparison of similarity learning methods

Methods	RI	Purity	NMI
Euclidean	0.4743	0.4633	0.0593
Cosine	0.4862	0.4654	0.0582
GMML	0.5024	0.4822	0.0698
LMNN	0.5778	0.5374	0.1148
K-means	0.6347	0.6659	0.2316
CNN_triple	0.7351	0.7561	0.3599

Table 8 The accuracy of methods in recommendation task

Methods	Accuracy
MLR	0.7086
Naive Bayes	0.7004
SVM	0.7238
CNN_ves	0.7857
SD_CNN_ves	0.8242

CNN_ves: It means that the vector representation is obtained through CNN, and then the result is obtained by the softmax function.

SD_CNN_ves (our model): This model firstly uses CNN to achieve similarity grouping, then uses CNN to obtain vector representation, and finally obtains the result through softmax function.

Table 8 shows the accuracy of methods in migration recommendation task. It can be concluded from Table 8 that the performance of the model proposed in this paper is better than other methods. It is because that it not only achieves similarity learning, but also considers the impact of historical medical visits and time information on future medical migration.

Figure 5 describes the ROC of several models. From Fig. 5, we can conclude that the area of ROC of the model proposed in this paper is the largest, that is, the AUC value is the largest. So the model in this paper has better performance.

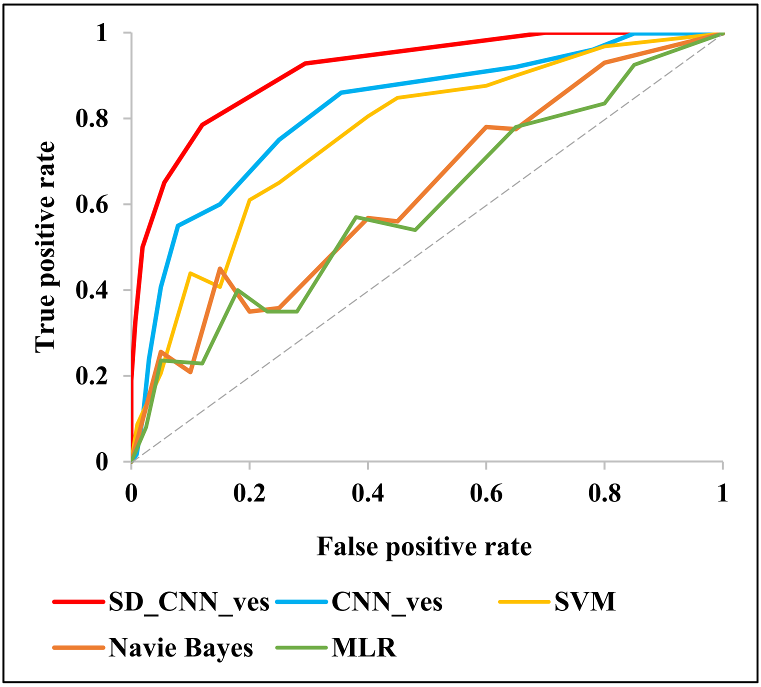

Fig. 5 The compassion of ROC

5 Conclusion

How to accurately predict the future medical treatment behaviors of patients from the historical health insurance data has become an important research issue in healthcare. In this paper, an Attention-based Bidirectional Gated Recurrent Unit (AB-GRU) medical treatment migration prediction model is proposed to predict which hospital patients will go to in the future. Due to medical treatment in different places has an important impact on the distribution of health insurance funds, when medical treatment prediction has been completed, this paper proposes a medical treatment migration recommendation model to recommend whether patients need medical treatment migration.

Acknowledgements This work was supported by the National Key Research and Development Plan of China (No.2018YFC0114709), the Natural Science Foundation of Shandong Province of China for Major Basic Research Projects (No.ZR2017ZB0419), the Taishan Industrial Experts Program of Shandong Province of China (No.tscy20150305).

References

1. Asfaw, L.S., Ayanto, S.Y., Aweke, Y.H.: Health-seeking behavior and associated factors among community in southern Ethiopia: community based cross-sectional study guided by health belief model. BioRxiv. 388769 (2018)
2. Baytas, I.M., Xiao, C., Zhang, X., Wang, F., Jain, A.K., Zhou, J.: Patient subtyping via time-aware lstm networks. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 65–74 (2017)
3. Bei, C., Juan, Y.Y.: Health seeking behavior of elderly floating population and the influence factors. Number 7. 856–859 (2015)
4. Bhojani, U., Beerenahalli, T.S., Devadasan, R., Munegowda, C.M., Devadasan, N., Criel, B., Kolsteren, P.: No longer diseases of the wealthy: prevalence and health-seeking for self-reported chronic conditions among urban poor in southern India. BMC Health Serv. Res. **13**(1), 306 (2013)
5. Che, C., Xiao, C., Liang, J., Jin, B., Zho, J., Wang, F.: An RNN architecture with dynamic temporal matching for personalized predictions of parkinson's disease. In: Proceedings of the 2017 SIAM International Conference on Data Mining, Houston, Texas, USA, April 27–29, 2017, pp. 198–206 (2017)
6. Che, Z., Yu, C., Zhai, S., Sun, Z., Liu, Y.: Boosting deep learning risk prediction with generative adversarial networks for electronic health records. In: 2017 IEEE International Conference on Data Mining, ICDM 2017, New Orleans, LA, USA, November 18–21, 2017, pp. 787–792 (2017)
7. Che, Z., Purushotham, S., Cho, K., Sontag, D., Yan, L.: Recurrent neural networks for multivariate time series with missing values. Sci. Rep. **8**(1), 6085 (2018)
8. Chen, J., Shao, J., He, C.: Movie fill in the blank by joint learning from video and text with adaptive temporal attention. Pattern Recogn. Lett. S0167865518302794 (2018)
9. Cheng, L., Ren, Y., Zhang, K., Shi, Y.: Medical treatment migration prediction in healthcare via attention-based bidirectional GRU. In: Web and Big Data - Third International Joint Conference, APWeb-WAIM 2019, Chengdu, China, August 1-3, 2019, Proceedings, Part I, pp. 19–34 (2019)
10. Choi, E., Bahadori, M.T., Song, L., Stewart, W.F., Sun, J.: GRAM: graph-based attention model for healthcare representation learning. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13–17, 2017, pp. 787–795 (2017)
11. Choi, E., Bahadori, M.T., Sun, J., Kulas, J., Schuetz, A., Stewart, W.: Retain: an interpretable predictive model for healthcare using reverse time attention mechanism. In: Advances in Neural Information Processing Systems, pp. 3504–3512 (2016)
12. Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y.: Attention-based models for speech recognition. In: Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7–12, 2015, Montreal, Quebec, Canada, pp. 577–585 (2015)

13. Duan, L.-z., Zhai, G.-q., Xuan, C.-y., Duan, G.-n., Zhang, Y., Geng, H.: The grey relational analysis of influential factors for chinese medicine in general hospital. In: Proceedings of 2011 IEEE International Conference on Grey Systems and Intelligent Services, pp. 23–29 (2011)
14. Hao, S., Sylvester, K.G., Ling, X.B., Shin, A.Y., Hu, Z., Jin, B., Zhu, C., Dai, D., Stearns, F., Widen, E., Culver, D.S., Alfreds, S.T., Rogow, T.: Risk prediction for future 6-month healthcare resource utilization in Maine. In: 2015 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2015, Washington, DC, USA, November 9–12, 2015, pp. 863–866 (2015)
15. He, D., Wang, S., Ruan, B., Zheng, B., Zhou, X.: Efficient and robust data augmentation for trajectory analytics: a similarity-based approach. World Wide Web. 1–27 (2019)
16. Kang, G., Ni, Z.: Research on early risk predictive model and discriminative feature selection of cancer based on real-world routine physical examination data. In: IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016, Shenzhen, China, December 15–18, 2016, pp. 1512–1519 (2016)
17. Kaur, S., Kalra, S.: Disease prediction using hybrid k-means and support vector machine. In: 2016 1st India International Conference on Information Processing (IICIP), pp. 1–6, 08 (2016)
18. Li, Q., Zhang, X., Xiong, J.J., Hwu, W.-m., Chen, D.: Implementing neural machine translation with bi-directional gru and attention mechanism on fpgas using hls. In: Proceedings of the 24th Asia and South Pacific Design Automation Conference, pp. 693–698 (2019)
19. Li, X., Zhou, Z., Chen, L., Gao, L.: Residual attention-based lstm for video captioning. World Wide Web. **22**(2), 621–636 (2019)
20. Xiao-jun, L.U., Zhang, N.: Study on influencing factors of the choice of hospitalization behaviors among agricultural transfer population. Chin. J. Health Policy. **11**(2), 10–16 (2018)
21. Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17–21, 2015, pp. 1412–1421 (2015)
22. Mohawish, A., Rathi, R., Abhishek, V., Lauritzen, T., Padman, R.: Predicting coronary heart disease risk using health risk assessment data. In: 17th International Conference on E-Health Networking, Application and Services, HealthCom 2015, Boston, MA, USA, October 14–17, 2015, pp. 91–96 (2015)
23. Ouyang, D., Zhang, Y., Shao, J.: Video-based person re-identification via spatio-temporal attentional and two-stream fusion convolutional networks. Pattern Recogn. Lett. **117**, 153–160 (2019)
24. Shang, J., Xiao, C., Ma, T., Li, H., Sun, J.: Gamenet: graph augmented memory networks for recommending medication combination. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 1126–1133 (2019)
25. Shao, F., Xian, Y.-T., Guo, J.-Y., Yu, Z.-T., Mao, C.-L.: A standard bibliography recommended method based on topic model and fusion of multi-feature. In: 2014 IEEE International Conference on Data Mining Workshop, pp. 198–204 (2014)
26. Suo, Q., Ma, F., Yuan, Y., Huai, M., Zhong, W., Gao, J., Zhang, A.: Deep patient similarity learning for personalized healthcare. IEEE Trans. Nanobiosci. (99), 1–1 (2018)
27. Suo, Q., Ma, F., Yuan, Y., Huai, M., Zhong, W., Zhang, A., Gao, J.: Personalized disease prediction using a cnn-based similarity learning method. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017, Kansas City, MO, USA, November 13–16, 2017, pp. 811–816 (2017)
28. Suo, Q., Xue, H., Gao, J., Zhang, A.: Risk factor analysis based on deep learning models. In: Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB 2016, Seattle, WA, USA, October 2–5, 2016, pp. 394–403 (2016)
29. Wang, L., Chen, J., Marathe, A.: A framework for discovering health disparities among cohorts in an influenza epidemic. World Wide Web. 1–24 (2018)
30. Wang, X., Wang, D., Xu, C., He, X., Cao, Y., Chua, T.-S.: Explainable reasoning over knowledge graphs for recommendation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 5329–5336 (2019)
31. Yayun, L.: Study on health-seeking behaviors of rural chronic patients in the view of the social determinants of health model. Med. Soc. (9), 14 (2015)
32. Yuan, Y., Xun, G., Jia, K., Zhang, A.: A multi-view deep learning method for epileptic seizure detection using short-time fourier transform. In: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB 2017, Boston, MA, USA, August 20–23, 2017, pp. 213–222 (2017)
33. Yuan, Y., Xun, G., Suo, Q., Jia, K., Zhang, A.: Wave2vec: learning deep representations for biosignals. In: 2017 IEEE International Conference on Data Mining, ICDM 2017, New Orleans, LA, USA, November 18–21, 2017, pp. 1159–1164 (2017)
34. Zeiler, M.D.: ADADELTA: an adaptive learning rate method. CoRR. abs/1212.5701 (2012)
35. Zhai, Y., Ge, W.U.: Analysis on medical behaviors of patients based on big data mining of electronic medical records (emr) information. J. Med. Inform. **38**(7), 12–17 (2017)

36. Zhan, M., Cao, S., Qian, B., Chang, S., Wei, J.: Low-rank sparse feature selection for patient similarity learning. In: 2016 IEEE 16th International Conference on Data Mining (ICDM), pp. 1335–1340 (2016)
37. Zhang, X.-w., Qiu, L.-j., Yang, Y.-n., Zhao, J., Fu, L.-c., Li, Q.-q.: Internet health information seeking behaviors of medical students under a medical internet perspective. Chin. Preventive Med. (3), 9 (2018)
38. Zheng, C., Zhu, J.: Grey relational analysis of factors affecting ipo pricing in China a-share market. In: 2017 International Conference on Grey Systems and Intelligent Services (GSIS), pp. 82–86 (2017)
39. Zheng, X., Ling, X.U.: Analysis of health seeking behavior based on the planned-action theory in rural area of China. Beijing da xue xue bao. Yi xue ban = Journal of Peking University. Health sciences. **42**(3), 270 (2010)
40. Zhu, Z., Yin, C., Qian, B., Cheng, Y., Wei, J., Wang, F.: Measuring patient similarities via a deep architecture with medical concept embedding. In: IEEE 16th International Conference on Data Mining, ICDM 2016, December 12–15, 2016, Barcelona, Spain, pp. 749–758 (2016)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Lin Cheng[1] • **Yuliang Shi**[1,2] • **Kun Zhang**[1,2]

[1] School of Software, Shandong University, Jinan, China

[2] Dareway Software Co., Ltd, Jinan, China