Dynamic prediction of cardiovascular disease using improved LSTM

Kuang Junwei, Hangzhou Yang, Liu Junjiang and Yan Zhijun
Beijing Institute of Technology, Beijing, China

Received 17 January 2019
Revised 24 February 2019
Accepted 24 February 2019

Abstract

Purpose – Previous dynamic prediction models rarely handle multi-period data with different intervals, and the large-scale patient hospital records are not effectively used to improve the prediction performance. This paper aims to focus on the prediction of cardiovascular disease using the improved long short-term memory (LSTM) model.

Design/methodology/approach – A new model based on the traditional LSTM was proposed to predict cardiovascular disease. The irregular time interval is smoothed to obtain the time parameter vector, and it is used as the input of the forgetting gate of LSTM to overcome the prediction obstacle caused by the irregular time interval.

Findings – The experimental results show that the dynamic prediction model proposed in this paper obtained a significant better classification performance compared with the traditional LSTM model.

Originality/value – In this paper, the authors improved the LSTM by smoothing the irregular time between different medical stages of the patient to obtain the temporal feature vector.

Keywords Cardiovascular disease, Dynamic prediction, LSTM

Paper type Technical paper

1. Introduction

Cardiovascular disease (CVD) is a class of diseases that involve the heart or blood vessels (Mendis *et al.*, 2011). CVD is the chronic disease that poses the greatest threat to people, and now it has become one of the leading causes of death around the world (Gibbons *et al.*, 1999). According to data released by the World Health Organization in May 2017, approximately 17.7 million people died of CVD in 2015, accounting for 31 per cent of the global death. Therefore, medical professionals and researchers have carried out extensive studies on the treatments and interventions for CVD (Kagashe *et al.*, 2017; Yan *et al.*, 2016; Zhu *et al.*, 2017).

Unlike acute diseases, there are many stages during the development and evolution of chronic diseases, and the characteristics of chronic diseases vary across stages (Asaria *et al.*, 2007). To take early interventions and maintain the healthiness of patients with CVD, various predictive models were developed to identify high-risk groups or predict the

International Journal of Crowd Science
Vol. 3 No. 1, 2019
pp. 14-25
Emerald Publishing Limited
2398-7294
DOI 10.1108/IJCS-01-2019-0002

This work was supported by National Key Research and Development Plan of China (Grant No: 2017YFB1400101), National Natural Science Foundation of China (Grant No: 71572013, 71872013, 71432002) and Beijing Municipal Social Science Foundation (Grant No: 18JDGLB040).

development of disease. Most previous disease prediction models were based on case-cohort study to investigate the relationship between potential high risk factors and morbidity and mortality (Ganna and Ingelsson, 2015). It is found that body mass index (Rost *et al.*, 2018), waist-hip ratio (Zwakenberg *et al.*, 2018) and sitting time and sitting posture (Howell *et al.*, 2017) have high correlations with the morbidity of CVD. However, due to the high cost of case-cohort study, the training data of these models are insufficient, and the prediction performance needs further improvement.

In recent years, with the development of information technology and the wide application of information systems in medical industry, hospital information system (HIS) has accumulated large-scale and multi-dimensional data including patient demographics, disease symptoms and diagnosis and biochemical indicators (Ahmadi *et al.*, 2017). HIS is an ideal data source to support risk assessment and the development of prediction models of CVD with machine learning algorithms (Goldschmidt, 2005). Long short-term memory (LSTM) is a recurrent neural network (RNN) that is suitable for processing and predicting important events with relatively long intervals and delays in time series. However, in the context of medical industry, the time interval between multiple hospitalizations of patients is different, and the traditional LSTM cannot effectively learn the important characteristics of patient's medical condition, which limits the practical application of LSTM in medical problems.

In this paper, we improved the LSTM by smoothing the irregular time between different medical stages of the patient to obtain the temporal feature vector. The temporal feature vector is used as the input of the forgetting threshold, which can effectively deal with the irregular time interval between the multi-period data and improve the predictive performance of the model.

2. Literature review

Machine learning algorithms have been used in various fields including disease prediction. Using logistic regression, Zhou *et al.* constructed a risk score model for type 2 diabetes in middle-aged male populations in rural China (Zhou *et al.*, 2017). Lin *et al.* considered the problem of co-occurring diseases and constructed a Bayesian multi-task learning model for chronic diseases and their corresponding complications (Lin *et al.*, 2017). To improve the accuracy of prediction, Long *et al.* proposed a hybrid heart disease prediction method, which combines rough set theory, clustering algorithm, genetic algorithm, naive Bayes and support vector machines, and the model showed obvious advantages over baseline models (Long *et al.*, 2015). Based on electronic medical record data, Ye *et al.* used the xgboost algorithm to predict the risk of hypertension of patients (Ye *et al.*, 2018).

However, existing models usually take a single period of sample data as input, and ignoring the time-series characteristics of clinical medical data, especially for chronic diseases. Therefore, many studies began to consider the inclusion of time series features in the model of chronic disease static prediction to construct a dynamic prediction model of chronic diseases. Marini *et al.* used the dynamic Bayesian model to simulate the long-term disease state of type 1 diabetes. The model can dynamically simulate the development of type 1 diabetes and predict future status (Marini *et al.*, 2015). Bueno *et al.* proposed to use a dynamic Bayesian network to model the patient's data over multiple periods to study the potential physiological changes that may occur after the patient received the drug treatment (Bueno *et al.*, 2016). Jackson *et al.* constructed a three-stage hidden Markov model (HMM) to characterize and predict chronic rejection after six months of lung transplantation (Jackson and Sharples, 2002). Forkan, *et al.* combined the HMM with neural network algorithms to learn and construct the probability of future disease in chronic diseases for elderly people living alone (Forkan and Khalil, 2017). However, both the dynamic Bayesian model and the HMM assume that the time interval between successive observations is fixed, and the

computational complexity increases rapidly as the number of variables increases, which limits the ability to learn complex data.

RNN is a kind of neural network used to process sequential data (Graves *et al.*, 2013b). The network memorizes the previous information and applies it to the calculation of the current output, that is, the nodes between the hidden layers of each segment also establish a connection. In addition, the input of the hidden layer at time step t includes the input of the input layer at time step t and the output of the hidden layer at time step t-1 (Graves, 2013a). However, in the process of learning long-term data, RNN may have the problem of gradient disappearing. In light of this, an improved version of RNN named LSTM was proposed to solve the problem of gradient disappearance in the long-dependent learning process by introducing structures of forgetting gates (Graves, 1997).

However, the existing LSTM model also assumes a fixed time interval between different time slices, which limits its practical application in medical problems. In view of the above problems, this paper improves the internal structure of the LSTM unit through parameterizes the time interval between time slices, thus obtaining important information of the influence of time interval on the development of disease.

3. Model framework

This paper investigates how to predict the diagnosis result at time step t (Y_T), in the case of a given patient's records from time step 1 to time step t ($X_1, X_2, \ldots, X_T$). Among them, the number of records per patient and the time interval between samples X_{t-1}, X_t, and X_{t+1} could be different.

To use LSTM to process sequence data with irregular time intervals, we first adapt the threshold structure of the LSTM unit to learn the temporal characteristics associated with CVD evolution at different time intervals. After that, we propose to use the target repeat prediction method for the output of hidden layer at each time step, which can simplify the model training process with different lengths of time series. Finally, for the output layer of the model, the Sigmoid function is introduced as the activation function of the multi-tag output, so that the patient's multiple diagnostic tags are predicted as output. The overall structure of the model is shown in Figure 1.

3.1 Introduction to long short-term memory

A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals, and the three gates regulate the flow of information into and out of the cell.

Figure 2 shows the structure of a traditional LSTM cell and illustrates the operations of the gates. There are three gates (input, forget and output) in the basic cell of LSTM, and each gate has a sigmoid activation function and a point-wise multiplication operation. The basic cell of the LSTM is defined as the following equations:

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \tag{1}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{2}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{3}$$

where f_t denotes the output of forget gate to the network at time step t, where σ is the logistic sigmoid function. i_t and o_t denote the output of input gate and output gate,

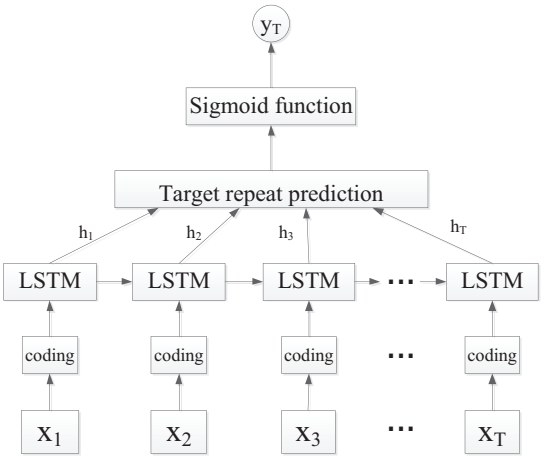

Figure 1.
Dynamic prediction model structure of cardiovascular disease

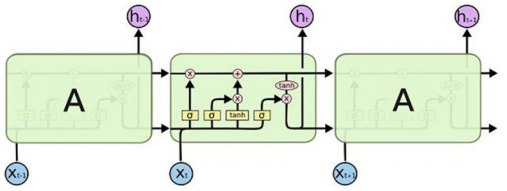

Figure 2.
Traditional LSTM

respectively. x_t and h_{t-1} are the input and the previous hidden state, respectively. W_f, W_i, W_o, b_f, b_i and b_o are weight matrices which are learned.

3.2 Improved long short-term memory

In the medical situation, patients with chronic diseases will go to the hospital because of the development of the disease, such as deterioration or recurrence. However, different patients may have different time intervals between hospitalizations due to their physical condition, condition, etc., and the difference may range from less than 1 month to several years. The lack of time interval brings certain difficulties and challenges to the study of clinical time series data.

To solve the problem of irregular time interval, we propose to smooth the time interval to obtain the time parameter vector and use it as the input of LSTM forget gate. The improved LSTM cell is shown in Figure 3. We will introduce the forward propagation process of the LSTM network.

The first step in the forward propagation of the LSTM network is the calculation of the forgotten threshold. This threshold determines which of the input information will be forgotten and will not affect future time step. In detail, the time interval between the time step t-1 and the time step t is smoothed to obtain a three-dimensional vector, and the time vector is used as an input parameter of the forget gate, as shown in equation (1).

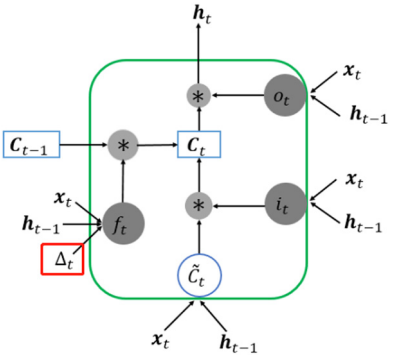

Figure 3.
Improved LSTM cell

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + P_f p_{\Delta_{t-1:t}} + b_f\right) \tag{4}$$

In equation (4), $P_f p_{\Delta_{t-1:t}}$ represents a vector after the smoothing of the time interval between time slices, and the smoothing formula is shown in equation (5):

$$p_{\Delta_{t-1:t}} = \left(\frac{\Delta_{t-1:t}}{60}, \left(\frac{\Delta_{t-1:t}}{180}\right)^2, \left(\frac{\Delta_{t-1:t}}{365}\right)^3\right) \tag{5}$$

In equation (5), $\Delta_{t-1:t}$ represents the time interval, in units of days. Because patients rarely re-hospitalize in the same month, so we choose two months as the denominator, then half a year and one year, making the vector $p_{\Delta_{t-1:t}}$ within a reasonable range.

P_f is a connection weight parameter corresponding to the time interval vector, which needs to be optimized for training to handle the memory effect generated by the irregular time interval.

The second step of forward propagation determines what information is saved in the cell state. First, you need to generate a temporary state and then update the old cell state. The formula is shown in equations (6) and (7).

$$\tilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C) \tag{6}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{7}$$

where W_C and b_C are the connection weight and offset of the temporary state. $\tilde{C}_t$ is a temporary state containing new candidate values. C_{t-1} is the status information of the previous time step. C_t is the state of the time step t after the update.

The third step of forward propagation determines the final network output, as shown in equation (8).

$$h_t = o_t * \tanh(C_t) \tag{8}$$

where h_t is the current hidden state, and h_t and C_t will be used as input for the next time step.

3.3 Target repeat prediction

When constructing a traditional LSTM network model, generally, only the output prediction of the last time step is given, and the error of the entire network is calculated to update the

network weight. However, when a sample has or is truncated into a short time series, the prediction performance could be worsened.

To solve the above problem, this paper adopts the target repeat prediction method. For the output of the hidden layer of each time step, the prediction probability of the diagnosis is calculated by the Sigmoid activation function, and the prediction loss of each time step is obtained by combining the real classification label. Finally, we use the weighted summation of the prediction loss of all time slices and the prediction loss of the last time slice as a loss function of the entire model to update the parameters of the entire model.

For a single time step, the loss function is calculated as follows:

$$\text{loss}(\hat{y}, \text{y}) = \frac{1}{|C|}\sum_{i=1}^{i=|C|} -\left(y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)\right) \tag{9}$$

In equation (9), $\hat{y}$ represents the disease diagnosis classification probability vector calculated by the Sigmoid function in a single time slice, and $\hat{y}_i$ represents the output probability of the corresponding i-th disease diagnosis. y indicates the actual class label of the current sample, and y_i indicates the classification label of the i-th disease diagnosis, taking 0 or 1. C represents the dimension of the classification label vector. The overall loss function for the entire model is shown in equation (10):

$$\text{loss} = \alpha \frac{1}{T}\sum_{t=1}^{T} loss\left(\hat{y}^{(t)}, \boldsymbol{y}^{(t)}\right) + (1 - \alpha) \cdot loss\left(\hat{y}^{T}, \boldsymbol{y}^{T}\right) \tag{10}$$

In equation (10), $\boldsymbol{y}^{(t)}$ is the real classification label of time slice t, and $\hat{y}^{(t)}$ represents the corresponding classification label prediction probability vector. α is the hyperparameter of the model, which is used to measure the sum of the predicted losses of all time slices and the weight of the last time slice prediction loss for the overall loss of the model.

3.4 Multi-label classification

In actual medical scenes, doctors make a disease diagnosis based on patient's laboratory indicators. Patients may have multiple diseases at the same time, such as coronary heart disease and type 2 diabetes. Thus, we define the disease diagnosis task as a multi-label classification task. This paper proposes a prediction model for multi-label classification, while the traditional model normally handles the single classification problem.

In the selection of the classification label, in addition to CVD, diseases that may cause CVD and diseases that may be caused by CVD are also included (Jonnagaddala *et al.*, 2015), such as hyperlipidemia, diabetes and so on, which can be divided into eight categories (c_1, c_2, ..., c_8). The diagnosis output for each sample is represented as eight-dimensional vectors with Boolean values. The i-th dimensional of the vector is 1 if the diagnosis belongs to c_i and 0 otherwise.

Compared to logistic and Softmax function, all elements in the output probability vector of Sigmoid function are not equal to 1, which is more suitable for multi-label classification problems. Therefore, we use Sigmoid as the activation function of our model. In existing multi-label classification studies, the classification result is the k-value element with the highest numerical value in the output probability vector, and the value of k is determined according to the actual problem (Tsoumakas *et al.*, 2007). In this paper, the average number of labels for all samples is about 3, and k is set to 3.

4. Experiment analysis

4.1 Data description

In the study, we used the data collected from the HIS of a hospital. The data set contained age, sex, 23 test indicators and nine disease diagnosis labels. The specific test indicators we used are shown in Table I. The disease diagnosis labels are shown in Table II. All information of patients that are recorded during hospitalization was identified by the patient ID and hospital ID. The disease diagnosis uses International Classification of Diseases 10th Revision (ICD-10) coding, and the test and inspection items use the system-defined code, which can be uniquely identified.

4.2 Data preprocess

To make the data meet the requirements and specifications, we only keep the records of patients whose "patient ID" and "hospital ID" is non-empty, the number of hospitalizations is more than twice, the "discharge method" is "normal", "age" is 18 or older, "admission time" and "discharge time" is valid, and the diagnostic records include cardiovascular or related diseases.

After preprocessing, we obtained 12,545 hospital records generated by 3,805 patients collecting from 15 March 1999 to 7 July 2010and calculated the length of time series for each sample. As shown in Figure 4, the length of samples is mostly concentrated from 2 to 5. Therefore, in the subsequent model training process, we set the maximum length of all samples to be 5. Samples with length less than 5 are complemented by 0, and samples with length longer than 5 are truncated.

The data set consists of 2,176 males and 1,629 females. The age distribution is shown in Table III.

Table I.
Test indicators

Column	Test indicator
1	Serum albumin (g/L)
2	Low density lipoprotein cholesterol (mmol/L)
3	Triglyceride (mmol/L)
4	High density lipoprotein cholesterol (mmol/L)
5	Glutamic-pyruvic transaminase (IU/L)
6	Glutamic oxalacetic transaminase (IU/L)
7	Creatine jubase (IU/L)
8	Creatine Kinase Isoenzyme (IU/L)
9	Creatinine (umol/L)
10	Kalium (mmol/L)
11	Alkaline phosphatase (mmol/L)
12	Fasting plasma glucose (IU/L)
13	Chlorine (mmol/L)
14	Sodium (mmol/L)
15	Urine-specific gravity (mmol/L)
16	urea nitrogen
17	Uric Acid (umol/L)
18	Urine ph
19	Globulin (g/L)
20	Lactic dehydrogenase (IU/L)
21	Direct bilirubin (umol/L)
22	Total cholesterol (mmol/L)
23	Total bilirubin (mg/dl)

The missing values of continuous variables are filled with mean values, and the missing values of discrete variables are filled with the majority.

To meet the input requirements of LSTM, we encode classification features using one-hot encoding. On the test project, the mean, maximum and minimum values of the sequence data are extracted to achieve feature extraction and dimensionality reduction.

Different variables have different value range and units, and the value range and unit dimensions have great impact on the weight learning process of the model. Generally, in classification and clustering algorithms, the z-score algorithm is usually used for normalization, which can achieve better results. Therefore, this paper uses the z-score standardization method to preprocess the input data.

4.3 Performance evaluation metrics

This paper focuses on the classification of disease diagnosis, and the classification performance of the diagnosis of different diseases. Therefore, the $Precision_{micro}$, $Recall_{micro}$

Column	Disease diagnosis label
1	Coronary heart disease
2	Myocardial infarction
3	Stenocardia
4	High blood pressure
5	Cerebral infarction
6	Hyperlipidaemia
7	Diabetes mellitus
8	Chronic renal insufficiency
9	Other

Table II.
Disease diagnosis labels

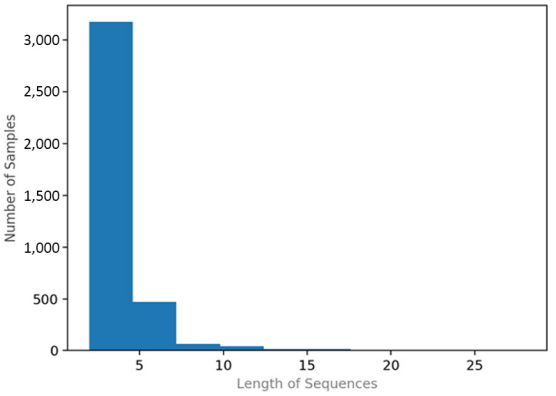

Figure 4.
Sample sequence length histogram

Min	1st	Median	Mean	3rd	Max
19	54	67	64	76	102

Table III.
Sample age distribution

and $F1_{micro}$ are selected as evaluation metrics. These three indicators are adapted from the corresponding single label classification model, and the calculation formulas are as follows:

$$Precision_{micro} = \frac{\sum_{j=1}^{q} TP_j}{\sum_{j=1}^{q} TP_j + \sum_{j=1}^{q} FP_j} \tag{11}$$

$$Recall_{micro} = \frac{\sum_{j=1}^{q} TP_j}{\sum_{j=1}^{q} TP_j + \sum_{j=1}^{q} FN_j} \tag{12}$$

$$F1_{micro} = \frac{2 * Precision_{micro} * Recall_{micro}}{Precision_{micro} + Recall_{micro}} \tag{13}$$

In addition, the AUC indicator indicates the area under the ROC curve and is often used to evaluate classifier performance. Therefore, in the multi-classification problem of this paper, we use micro AUC as one of the model evaluation indicators.

4.4 Experimental result

LSTM learn the characteristics of data set from training set and predict the classification labels of new samples. The hyper parameters of LSTM model needs to be set. The proposed improved LSTM model is defined as T-LSTM-TR. We train and tune the parameters of our model using 10-fold cross-validation method.

The hyper parameters need to be adjusted and optimized during training process, including the number of hidden layer neurons H, the end time slice loss function weight α and the dropout parameter. The model is trained by setting different parameter sets separately, and then the test results are compared. Finally, the optimal parameters of the T-LSTM-TR model is set as H = 120, $\alpha = 0.5$ and Dropout = 0.4.

This paper selects the traditional LSTM model as the benchmark model for performance comparison. As shown in Table IV, the performance of T-LSTM-TR model proposed in this paper is similar to that of the LSTM model in terms of precision, while the performance of T-LSTM-TR is significantly superior compared to that of the traditional LSTM model in terms of other indicators. The results show that the classification performance of our model is effectively improved by adapting the departmental structure of traditional LSTM unit. As shown in Figure 5, we can more clearly compare the performance of T-LSTM-TR and LSTM through the ROC curve.

For the hidden layer feature processing of all time slices, the average pooling process is an alternative method, and the output prediction result can be obtained using the Sigmoid function. To validate the effectiveness of the proposed target repeat prediction method, we

Table IV. The performances of T-LSTM-TR and LSTM

Models	Precision	Recall	F1	AUC
T-LSTM-TR	*0.492*	*0.811***	*0.608**	*0.896****
LSTM	0.478	0.754	0.584	0.844

Notes: *$p < 0.1$; **$p < 0.05$; ***$p < 0.001$

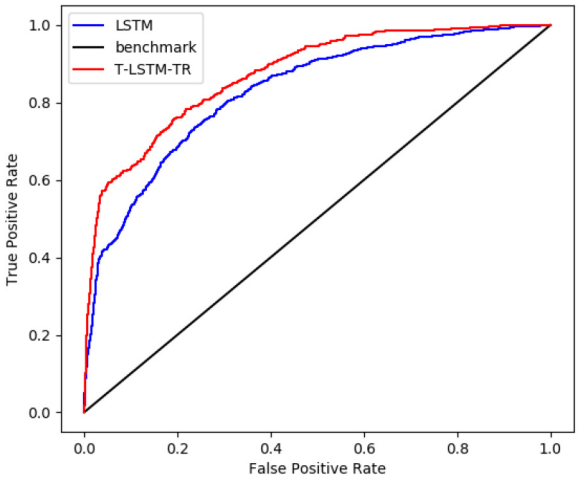

Figure 5.
ROC curve of T-LSTM-TR and LSTM

compared the performance of average pooling process and target repeat prediction method. The average pooled model is defined as T-LSTM-MP, and the comparison results are shown in Table V. As shown in Table V, the T-LSTM-TR model obtained higher results compared to the T-LSTM-MP model in all the four indicators, indicating that the target repeated prediction method is significantly better than the average pooling method. As shown in Figure 5, we can more clearly compare the performance of T-LSTM - TR and T-LSTM-MP through the ROC curve.

5. Conclusion

Based on the traditional LSTM, this paper proposed a new model by improving the internal forgetting gate input. First, the irregular time interval is smoothed to obtain the time parameter vector, and then it is used as the input of the forgetting gate to overcome the prediction obstacle caused by the irregular time interval. The experimental results show that the dynamic prediction model proposed in this paper has a significant improvement in classification performance compared with the traditional LSTM model, which verifies the effectiveness of the proposed model.

There are still some limitations in this paper for future studies. First, this paper assumes that the diagnostic labels of the samples are independent to each other, which in fact there are varying degrees of correlation between many diseases. Second, due to the limits of data size, although the model of this paper has a significant improvement over the existing models in the performance evaluation indicators, the model still need further improvement to meet the requirements of practical applications.

模型	Precision	Recall	F1	AUC
T-LSTM-TR	*0.492****	*0.811****	*0.608****	*0.896****
T-LSTM-MP	0.478	0.787	0.591	0.879

Notes: *$p < 0.1$; **$p < 0.05$; ***$p < 0.001$

Table V.
The performance of T-LSTM-TR and T-LSTM-MP

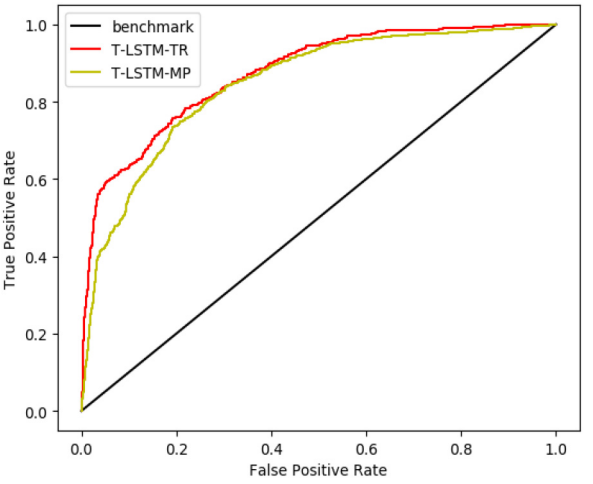

Figure 6.
ROC curve of T-LSTM-TR and T-LSTM-MP

References

Ahmadi, H., *et al.* (2017), "Hospital information system adoption: expert perspectives on an adoption framework for Malaysian public hospitals", *Computers in Human Behavior*, Vol. 67, pp. 161-189.

Asaria, P., *et al.* (2007), "Chronic disease prevention: health effects and financial costs of strategies to reduce salt intake and control tobacco use", *Lancet*, Vol. 370 No. 9604, pp. 2044-2053.

Bueno, M.L., *et al.* (2016), "Understanding disease processes by partitioned dynamic bayesian networks", *Journal of Biomedical Informatics*, Vol. 61, pp. 283-297.

Forkan, A.R.M. and Khalil, I. (2017), "PEACE-Home: Probabilistic estimation of abnormal clinical events using vital sign correlations for reliable home-based monitoring", *Pervasive and Mobile Computing*, Vol. 38 No. 2, pp. 296-311.

Ganna, A. and Ingelsson, E. (2015), "5 year mortality predictors in 498103 UK biobank participants: a prospective population-based study", *Lancet*, Vol. 386 No. 9993, pp. 533-540.

Gibbons, R.J., *et al.* (1999), "ACC/AHA/ACP-ASIM guidelines for the management of patients with chronic stable angina: a report of the American college of cardiology/american heart association task force on practice guidelines (committee on management of patients with chronic stable angi", *Journal of the American College of Cardiology*, Vol. 99 No. 1, pp. 2092-2197.

Graves, A. (1997), "Long short-term memory", *Neural Computation*, Vol. 9 No. 8, pp. 1735-1780.

Graves, A. (2013a), *Generating Sequences with Recurrent Neural Networks*, *Computer Science*.

Graves, A., *et al.* (2013b), *Speech Recognition with Deep Recurrent Neural Networks*, Vol. 38 No. 2003, pp. 6645-6649.

Howell, C.R., *et al.* (2017), "Clinical impact of sedentary behaviors in adult survivors of acute lymphoblastic leukemia: a report from the St. Jude lifetime cohort study", *Cancer*, Vol. 124 No. 5, pp. 1036-1043.

Jackson, C.H. and Sharples, L.D. (2002), "Hidden markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients", *Statistics in Medicine*, Vol. 21 No. 1, pp. 113-128.

Jonnagaddala, J., *et al.* (2015), "Identification and progression of heart disease risk factors in diabetic patients from longitudinal electronic health records", *Biomed Research International*, Vol. 2015, pp. 1-10.

Kagashe, I., *et al.* (2017), "Enhancing seasonal influenza surveillance: topic analysis of widely used medicinal drugs using twitter data", *Journal of Medical Internet Research*, Vol. 19 No. 9, pp. e315.

Lin, Y.K., *et al.* (2017), "Healthcare predictive analytics for risk profiling in chronic care: a bayesian multitask learning approach", *Mis Quarterly*, Vol. 41 No. 2, pp. 473-495.

Long, N.C., *et al.* (2015), "A highly accurate firefly based algorithm for heart disease prediction", *Expert Systems with Applications*, Vol. 42 No. 21, pp. 8221-8231.

Marini, S., *et al.* (2015), "A dynamic Bayesian network model for long-term simulation of clinical complications in type 1 diabetes", *Journal of Biomedical Informatics*, Vol. 57 No. C, pp. 369-376.

Mendis, S., *et al.* (2011), *Global Atlas on Cardiovascular Disease Prevention and Control*, *Geneva World Health Organization*.

Rost, S., *et al.* (2018), "New indexes of body fat distribution and sex-specific risk of total and cause-specific mortality: a prospective cohort study", *Bmc Public Health*, Vol. 18 No. 1, pp. 427.

Tsoumakas, G., *et al.* (2007), "Multi-label classification: an overview", *International Journal of Data Warehousing and Mining*, Vol. 3 No. 3, pp. 1-13.

Yan, Z., *et al.* (2016), "Knowledge sharing in online health communities: a social exchange theory perspective", *Information and Management*, Vol. 53 No. 5, pp. 643-653.

Ye, C., *et al.* (2018), "Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning", *Journal of Medical Internet Research*, Vol. 20 No. 1, p. e22.

Zhou, H., *et al.* (2017), "Development and evaluation of a risk score for type 2 diabetes mellitus among middle-aged Chinese rural population based on the RuralDiab study", *Scientific Reports*, Vol. 7, p. 42685.

Zhu, L., *et al.* (2017), "Mining medical related temporal information from patients' self-description", *International Journal of Crowd Science*, Vol. 1 No. 2, pp. 110-120.

Zwakenberg, S.R., *et al.* (2018), "Bone markers and cardiovascular risk in type 2 diabetes patients", *Cardiovascular Diabetology*, Vol. 17, p. 45.

Corresponding author
Yan Zhijun can be contacted at: yanzhijun@bit.edu.cn