Chinese Journal of Electronics
Vol.29, No.3, May 2020

A Feature Optimized Deep Learning Model for Clinical Data Mining

WU Tianshu[1], CHEN Shuyu[1], TIAN Yingming[2] and WU Peng[2]

(1. *College of Computer Science, Chongqing University, Chongqing 400044, China*)

(2. *College of Automation, Chongqing University, Chongqing 400044, China*)

Abstract — the Artificial intelligence (AI) has gradually changed from frontier technology to practical application with the continuous progress of deep learning technology in recent years. In this paper, the Random forest (RF) algorithm is adopted to preprocess and optimize the feature subset of ICU data sets. Then these optimized feature subsets are used as input of Long shortterm memory (LSTM) deep learning model, and the early disease prediction of ICU inpatients is carried out by the method of neural network deep learning. Experiments show that this prediction method has higher prediction accuracy compared with other machine learning and deep learning models.

Key words — Clinical data mining, Random forests (RF), Long short-term memory (LSTM).

I. Introduction

Through the application of Artificial intelligence (AI) technology, the electronic health data of each patient can be excavated and analyzed to help doctors improve the accuracy of medical diagnosis, discover patient's early diseases and predict the development trend of diseases. It has important application value for optimizing the process of diagnosis and treatment, improving the efficiency of diagnosis and treatment, and reducing medical expenses[1].

The traditional statistical analysis methods have some shortcomings, such as low efficiency and accuracy, in the face of high-dimensional and large-scale Intensive care database. Machine learning, including semiotic learning represented by decision tree model[2], connectionist learning represented by Artificial neural network (ANN) model[3, 4], and statistical learning represented by Support vector machine (SVM) model[5] has gradually led the research of AI. AI algorithm can use a large number of sample data for data training and knowledge extraction, analyze the actual application effect of the model, and make corresponding feedback and adjustment to the model. Therefore, it is more accurate in disease detection and prediction.

Some researchers have done research on disease severity prediction based on Electronic medical record (EMR). These studies show that the physiological indicators of ICU patients can be mined and analyzed to predict the mortality, length of stay and diseases classification after admission, optimize the allocation of medical resources, and intervene the severe patients as early as possible, thus to save the lives of patients. These studies mostly use simple logistic regression model[6], which is easy to be under fitted, resulting in insufficient classification accuracy;prediction is often applicable to a single disease type[7]; or simple integration of multiple disease type prediction models[8], resulting in the model system is too complex, the identified features is difficult to effectively evaluate the severity of patients.

This paper studies the application of Long shortterm memory (LSTM) deep learning model based on the feature optimized by Random forest (RF) algorithm in clinical data mining. Experiments show that it has higher prediction accuracy than other machine learning and deep learning models.

II. Feature Extraction Based on RF Algorithm

There are many problems in clinical data due to noise, instrument failure and human error, such as a large number of erroneous items (as well as outliers and missing

Manuscript Received Sept. 25, 2019; Accepted Dec. 9, 2019. This work is supported by National Natural Science Foundation of China (No.61272399, No.61572090), University Doctoral Program Fund of the Ministry of Education (No.20110191110038), and the Chongqing Population Health Big Data Engineering Technology Research Center (No.cstc2016yfpt_gcjsyjzx0189)

DOI:10.1049/cje.2020.03.004

values), inconsistent data standards, low data availability *etc.*[9]. The time series characteristics of the data also make the observation results affected by irregularity, uneven interval and time constraints. Direct use of the original data will lead to deviations in the statistical analysis results of clinical trials, and the credibility of the conclusions will be significantly affected. Therefore, it is necessary to pre-process these data before mining in order to meet the needs of data mining algorithms and improve data quality and analysis results[10].

RF is an algorithm that integrates multiple decision trees and makes decisions by voting on multiple decision trees[11]. In the decision tree, random processes are added to the decision tree in the direction of row and column respectively. When building decision tree in row direction, bootstrapping (playback sampling) is used to get training data, and non-return random sampling is used to get feature subset in column direction. Based on this, the optimal segmentation point is obtained, which is the basic principle of RF algorithm, as shown in Fig.1.

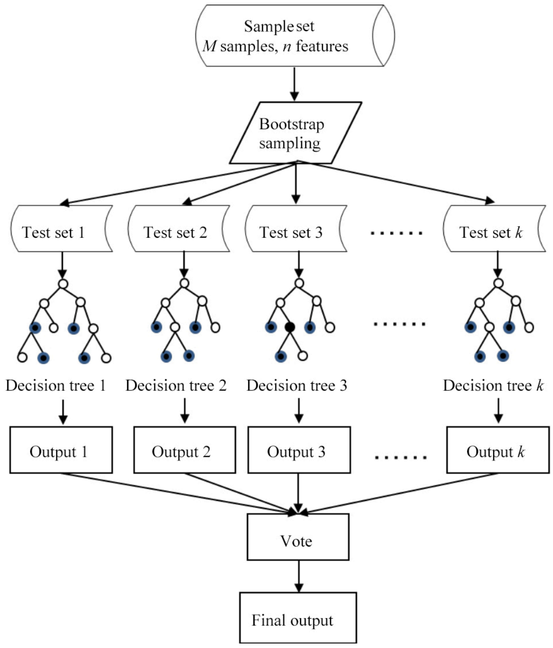

Fig. 1. Principle of RF algorithms

In this paper, RF algorithm is used to extract features from sample data. Using bootstrapping method, n training samples are randomly sampled from the original training sample set, and a new training sample set is generated. A total of k-round extraction was carried out to obtain k training sets.

For k training sets, we train k models and get a classification model sequence $\{h_1(x), h_2(x), \cdots, h_k(x)\}$. A "vote" will be held accordingly. Then a multiclassification model system is constructed by using the sequence of classification models. The final classification results are obtained by simple majority voting method. The output results is shown in Eq.(1).

$$H(x) = \arg\max_{Y} \sum_{i=1}^{k} I\left(h_i(x) = Y\right) \tag{1}$$

where $H(x)$ is a combination classification model, h_i is a single decision tree classification model, Y is an output variable, and $I(\cdot)$ is an indicator function of the set.

RF increase the difference between classification models by constructing different training sets, and improve the extrapolation and prediction ability of combined classification models. When classifying a new object based on certain attributes, each tree in a RF will give its own classification choice. For the classification problem, the output of the forest will be the most voted classification option;for the regression problem, the mean of the predicted results of k models will be taken as the final prediction result.

We use the R^2 formula in reference[12] to evaluate the fitting effect between the predicted value and the actual value of a single feature, as shown in Eq.(2). Finally, each feature is ranked according to its score. The higher the score, the better the prediction performance of the feature, but the maximum score is not more than 1.

$$R^2 = 1 - \frac{\sum_{i-1}^{n} \left(y_i - \hat{y}_i\right)^2}{\sum_{i-1}^{n} \left(y_i - \bar{y}_i\right)^2} \tag{2}$$

where

y_i: the true value of sample response;

$\hat{y}i$: the predicted value of a single feature in the model;

$\bar{y}i$: the average of the true value.

Compared with other algorithms, the RF algorithm can effectively improve the classification accuracy of new samples, and have strong nonlinear fitting ability in the feature space. Because the nodes of the decision tree can be randomly selected to divide the features, the model can still be effectively trained when the feature dimension of the sample is very high.

The training of RF model is highly parallel and has the advantage of large sample training speed.

RF has strong anti-interference ability. For unbalanced data sets, RF can provide an effective way to balance the error of data sets. When a large amount of data is lost, RF algorithm can still maintain the accuracy.

RF can solve the problems of classification and regression, and have good estimation performance in both aspects.

III. LSTM Deep Learning Model

Clinical medical data is large in scale, fast in updating, and has the characteristics of multi-modality, incompleteness, timing, redundancy, privacy *etc*. Using deep learning algorithm to analyze various medical information in clinical data can more accurately mine valuable predictive results and help doctors diagnose[13].

Typical deep learning models include Deep feed-forward network (DFN), Convolutional neural network (CNN), Deep belief network (DBN) and Recurrent neural network (RNN).

DFN also known as Feedforward neural network (FNN) or Multilayer perceptron (MLP), is a typical deep learning model. The whole process from input to output through intermediate calculation is feed forward, and there is no feedback connection.

The training of neural networks usually uses iterative gradient-based optimization to make the cost function reach a very small value. Instead of guaranteeing global convergence as a linear equation solver for training linear regression models or a convex optimization algorithm for training logical regression or SVM.

CNN is a kind of FNN which contains convolution computation and has deep structure. It is one of the representative algorithms of deep learning and is often used to analyze visual images[14].

DBN consists of several Restricted Boltzmann machines (RBM) layers. These networks are “restricted” to a visual layer and a hidden layer. There are connections between layers, but there are no connections between cells in the layer. Hidden layer units are trained to capture the correlation of higher-order data in the visual layer[15].

RNN is the only algorithm with memory, it can remember the important information received and form a profound understanding of the sequence and its context. This enables it to generate predictions very accurately in continuous data. Therefore, they are the preferred algorithm for time series data.

The structure of RNN is to connect the edge of the hidden layer to the next hidden layer. Unlike traditional MLP, it is related to time. Next time will be affected by this time. To better illustrate this point, we can expand the network according to time, as shown in Fig.2.

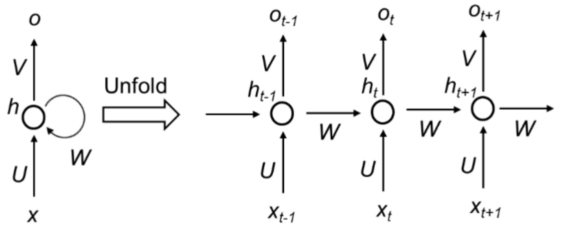

Fig. 2. Expansion diagram of RNN by time

The forward propagation of RNN is calculated by time series. The accumulated residuals propagate back from the last time.

h_t is the hidden state at time t, which is calculated by the state h_{t-1} of the previous time and the input x_t of the current time plus the weight, as Eq.(3).

$$h_t = \sigma\left(Ux_t + Wh_{t-1} + b\right) \tag{3}$$

where U and W are weight coefficient matrices, b are biased vectors of each layer, and σ is the activation function.

In the case of long sequence, RNN often encounters the problem of gradient explosion or disappearance because of the long backpropagation path. To this end, some researchers have proposed an improved LSTM neural network method[16]. The core of the improved method is to construct a gate unit with parameters, thus the model can selectively remember those important information.

LSTM can remember input for a long time because it contains information in memory and can read, write and delete information from memory. LSTM network helps to solve the problem of gradient memory loss, thus replacing the relatively insensitive learning method.

LSTM network consists of a basic unit called Cell. Cell’s function is to form memory at any time interval, thus forming a large number of memory networks. The neural network consists of three gates: input gate i_t, forgetting gate f_t and output gate o_t. Each door can be regarded as a traditional artificial neuron for calculating the activation of the weighted sum. The core of LSTM network is Block, whose structure is shown in Fig. 3.

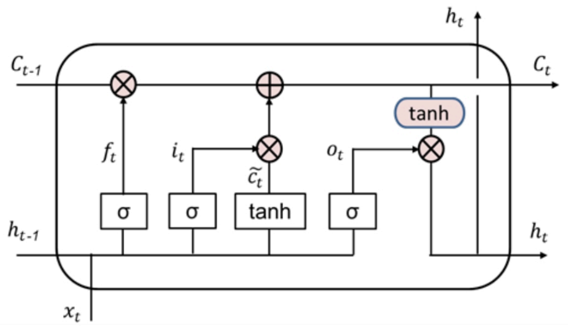

Fig. 3. Structure of LSTM block

Block states are similar to conveyor belts, running directly across the chain, with only a few linear interactions. Cell remembers the dependencies between elements in the input sequence, and the three gates control the flow of information into and out of the block.

f_t is the forget gate, which indicates whether the historical information stored by the current hidden layer node is retained. The gate reads h_{t-1} and x_t, and outputs a value between 0 and 1 to each number in the cell state C_{t-1}. 1 means “complete retained” and 0 means “complete abandonment”. The calculation method is as Eq.(4).

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \tag{4}$$

i_t is the input gate, which decides how much new information is added to the current hidden layer node and outputs a value between 0 and 1 to each number in the cell state C_{t-1}. 1 means “complete retained” and 0 means “complete abandonment”. The calculation method is as Eq.(5).

$$i_t = \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right) \tag{5}$$

c_t is the final internal memory of unit. It consists of the dot product of the memory c_{t-1} of the forget gate and the last location plus the dot product of the input gate and the candidate state $\widetilde{C}_t$. The calculation method is as Eq.(6).

$$c_t = f_t \odot c_{t-1} + i_t \odot \widetilde{c}_t \tag{6}$$

where $\widetilde{c}_t = \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right)$

O_t is the output gate. If it is 1, it means that the output value of the current node is output to the next layer. The calculation method is as Eq.(7).

$$o_t = \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \tag{7}$$

h_t is the hidden state. The calculation method is as Eq.(8).

$$h_t = o_t \odot \tanh\left(c_t\right) \tag{8}$$

The overall framework of LSTM prediction model is shown in Fig.4.

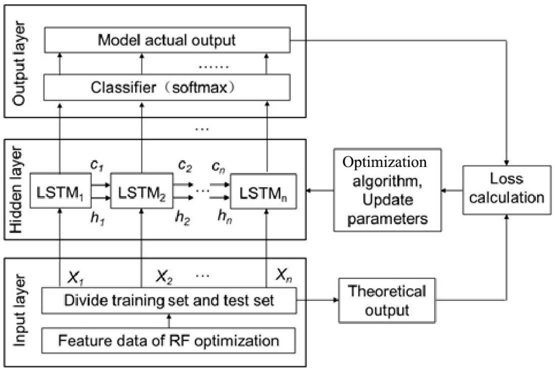

Fig. 4. Overall framework of LSTM prediction model

The feature data optimized by RF algorithm will be input into the input layer of LSTM. 70% of the data were selected as training set and 30% as test data. Data is divided into $(x_1, x_2, \cdots, x_n)$ by data segmentation. Then the data set is input into the hidden layer, and the data passing through the hidden layer will reach the classifier of the output layer of the model, and then the actual type output can be obtained. There is often a big gap between the real value and the theoretical value of the single training output, therefore we need to select a loss function to judge the error between the real value and the theoretical value.

We treat early clinical warning as a classification problem. Generally, cross entropy is used as loss function for classification problems. In the multi classification problem, for the samples (x, y), y is the real tag, and the prediction tag is the set of all tags. We assume that there are k tag values, The probability that the i-th sample is predicted to be the k-th tag is $p_{i,k}$, and there are n samples in total, then the total loss function of the data set is as Eq.(9).

$$\text{loss} = -\sum_{k=1}^{M} y_{i,k} \log p_{i,k} \tag{9}$$

where M is number of classes, $y_{i,k}$ is binary indicator (0 or 1) indicating whether the class label k is the correct classification for observation i. $p_{i,k}$ is the predicted output value, $y_{i,k}$ is the actual output value.

The objective of optimization is to minimize the loss function. Given the random seed number and learning rate of network initialization, the optimization algorithm is applied to update the network parameters continuously, and then the final hidden layer network is obtained.

IV. Experimental Analysis

This study used the open ICU clinical data set MIMIC III[17]. The data information included the patient’s demographic characteristics, vital signs, laboratory test results, in-patient circulation, treatment process, drug use, fluid inflow and outflow, nursing log, image report and discharge records.

A separate sample is created for each patient through patient table in data set. Each sample includes not only non-temporal features such as patient’s age, gender, type of admission, disease coding, but also output events, chart events and laboratory events were constructed as time series records of hospitalization events.

1. Experimental method

In this study, test data were used to predict the disease classification of patients after admission, and four different machine learning and deep learning methods were compared for analysis.

70% of the samples were selected as training samples and 30% as testing samples. Through 20 bootstrap sampling of the samples, the parameters were selected, and the average values of accuracy, sensitivity and specificity were obtained. Referring to the experimental method of reference[18], the accuracy, sensitivity and specificity were calculated by the following formulas as shown in Eqs.(10)–(12) respectively.

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \quad (10)$$

$$Sensitivity = TP/(TP + FN) \quad (11)$$

$$Specificity = TN/(TN + FP) \quad (12)$$

where TP means that positive samples are classified correctly; TN means that negative samples are classified correctly; FN means that negative samples are classified incorrectly; FP means that positive samples are classified incorrectly.

Accuracy refers to the degree of consistency between the measured results and the true value; *Sensitivity* refers to the probability of no missed diagnosis in predicting the condition; *Specificity* refers to the probability of no misdiagnosis in predicting the condition.

The experimental results were evaluated by confusion matrix. The confusion matrix is a matrix representation of the classification results. The confusion matrix is shown in Table 1.

Table 1. Confusion matrix

Observed value Predicted value	0	1
0	TP	FP
1	FN	TN

2. Experimental result

We take the prediction of mortality, length of stay and disease classification as the research content.

Mortality prediction: we regard the mortality prediction after admission as a multi classification problem, and predict the death of patients within 24 hours, 48 hours, 72 hours, 30 days and the whole hospitalization cycle respectively. The patients who died in the hospitalization cycle are marked as positive samples, while the other patients are marked as negative samples.

Hospitalization time prediction: in machine learning, hospitalization time prediction is usually treated as a regression problem, but in reality, hospitals only need to predict a rough range for planning treatment plans. Therefore, we divide the inpatient data into 10 categories: less than 1 day, 2 days, 3 days, 4 days, 5 days, 6 days, 7 days, 8 days, 8-14 days and more than 14 days, The prediction of inpatient time has changed from regression problem to multi classification problem.

Disease classification: we determine the patient's disease classification according to the International classification of diseases (ICD) diagnostic code group. Almost every health condition can be assigned a unique ICD-9 code. According to the single level classification system provided by the clinical classification software, we define the symptoms, use the standard ICD-9 coding group as the basis[19], remove the original part of redundancy and noise, and divide the coding of similar diseases into the same classification.

In this study, test data were used to predict the disease classification of patients after admission, and four different machine learning and deep learning methods were compared for analysis. According to the experimental steps, LSTM model (with RF algorithm optimized), RNN, DBN and SVM were used to calculate the experimental results. The accuracy, sensitivity and specificity of different models are shown in Table 2.

Table 2. Comparison of different models

Item	Optimized LSTM	RNN	DBN	SVM
Accuracy(%)	92.46	85.83	82.27	85.06
Sensitivity(%)	90.52	87.27	86.13	88.13
Specificity(%)	93.18	89.74	86.49	78.49

Table 2 shows that although the application of SVM which belongs to traditional machine learning is mature, its accuracy, sensitivity and specificity are obviously inferior to those of DBN, RNN and LSTM. At the same time, LSTM and RNN networks based on time series are better than DBN in disease prediction because of the existence of time sequence related data records in ICU data sets. LSTM model is outstanding in clinical data processing of time series, it can capture long-term dependence and nonlinear dynamic changes. The LSTM based on RF optimization has higher accuracy, sensitivity and specificity than RNN.

The graphical representation of sensitivity, specificity and accuracy curves in four different models are shown in Fig.5.

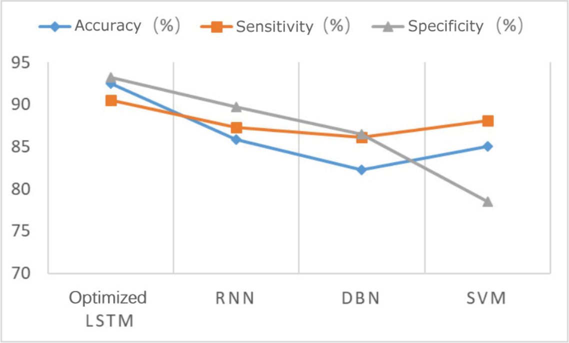

Fig. 5. Graphical representation in different models

V. Conclusions

SVM and shallow neural network are inferior to deep learning model in mining and learning deep data. RNN can remember the received important input information and form a profound understanding of the sequence and its context. This enables them to generate predictions very accurately in continuous data. Therefore, they are the preferred algorithm for time series data. LSTM has the strongest disease judgment ability. Although the structure of LSTM neural network is complex, its ability to judge diseases is superior to other methods. Even after multi-level learning, it can maintain memory without memory failure.

The improved LSTM network achieves higher accuracy, sensitivity and specificity than SVM, RNN and DFN in early prediction of inpatient diseases. Using LSTM deep learning model to mine massive medical data can diagnose diseases more accurately, reduce medical costs and improve medical level and efficiency.

References

[1] D. W. Bates, S. Saria, L. O. Machado, *et al.*, "Big data in health care: Using analytics to identify and manage high-risk and high-cost patients", *Health Affairs*, Vol.33, No.7, pp.1123–1131, 2014.

[2] C. H. Zhou Q. Wang, K. Z. Wu, *et al.*, "Averaged one-dependence decision trees ensemble algorithm", *Acta Electronica Sinica*, Vol.38, No.2, pp.434–438, 2010. (in Chinese)

[3] G. Falavigna, G. Costantino, R. Furlan, *et al.*, "Artificial neural networks and risk stratification in emergency departments", *Internal and Emergency Medicine*, Vol.14, No.2, pp.291–299, 2019.

[4] X. G. Gao, F. Li and K. F. Wan, "Accelerated learning for restricted boltzmann machine with a novel momentum algorithm", *Chinese Journal of Electronics*, Vol.27, No.3, pp.483–487, 2018.

[5] J. Shu, S. Liu, L. L. Liu, *et al.*, "Research on link quality estimation mechanism for wireless sensor networks based on support vector machine", *Chinese Journal of Electronics*, Vol.26, No.2, pp.377–384, 2017.

[6] D. M. Mu and K. Ren, "Comparison of three data mining algorithms in knowledge discovery of EMR", *Modern Library and Information Technology*, Vol.32, No.6, pp.102–109, 2016. (in Chinese)

[7] R. Duggal, S. Shukla, S. Chandra, *et al.*, "Impact of selected pre-processing techniques on prediction of risk of early readmission for diabetic patients in India", *International Journal of Diabetes in Developing Countries*, Vol.36, No.4, pp.469–476. 2016.

[8] K. S. Li, J. P. Fan, F. F. Zhou, *et al.*, "Early diagnosis algorithms for ICU emergencies", *Integration Technology*, Vol.1, No.2, pp.13–19, 2012. (in Chinese)

[9] N. Lavrac and B. *Zupan, Data Mining in Medicine*, Springer, New York, USA, pp.1111–1136, 2010.

[10] U. R. Acharya, S. L. Fernandes, J. E. WeiKoh, *et al.*, "Automated detection of alzheimer's disease using brain MRI images-A study with various feature extraction techniques", *Journal of Medical Systems,* Vol.43, No.9, pp.1–14, 2019.

[11] M. Khalilia, S. Chakraborty and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest", *BMC Medical Informatics and Decision Making*, Vol.11, No.1, pp.51–51, 2011.

[12] S. P. Zeng and J. L. Wang, "Predicted molecular ratio of aluminum reduction based on random forest and neural network", *Light Metals*, No.12, pp.21–25. 2018. (in Chinese)

[13] G. Litjens, T. Kooi, B. E. Bejnordi, *et al.*, "A survey on deep learning in medical image analysis", *Medical Image Analysis*, Vol.42, No.9, pp.60–88, 2017.

[14] E. Lima, X. Sun, J. Dong, *et al.*, "Learning and transferring convolutional neural network knowledge to ocean front recognition", *IEEE Geoscience and Remote Sensing Letters*, Vol.14, No.3, pp.354–358, 2017.

[15] F. Ghasemi, A. Fassihi, H. P. Sanchez, *et al.*, "The role of different sampling methods in improving biological activity prediction using deep belief network", *Journal of Computational Chemistry*, Vol.38, No.4, pp.195–203. 2017.

[16] Y. S. Zhang, J. Zheng, Y. L. Jia, *et al.*, "A text sentiment classification modeling method based on coordinated CNN-LSTM-attention model", *Chinese Journal of Electronics*, Vol.28, No.1, pp.120–126, 2019.

[17] A. E. Johnson, T. J. Pollard, L. Shen, *et al.* , "MIMIC-III, a freely accessible critical care database", *Scientific Data*, Vol.3, No.160035, DOI:10.1038/sdata.2016.35, 2016.

[18] M. Kumari and S. Godara, "Comparative study of data mining classification methods in cardiovascular disease prediction", *International Journal of Computer Science and Technology*, Vol.2, No.2, pp.304–308, 2011.

[19] International Classification of Diseases, Ninth Revision ICD-9 PDF. Available at:*https://simba.isr.umich.edu/restricted/docs/Mortality/icd_0 9_codes.pdf.*

WU Tianshu was born in 1989. He received the B.S. degree from Chongqing University of Post and Telecommunication. He is now a Ph.D. candidate in College of Computer Science of Chongqing University. His research interests include cloud computing, distributed computing and data mining. (Email: pwu@cqu.edu.cn)

CHEN Shuyu (corresponding author) was born in 1963. He received the B.S., M.S. and Ph.D. degrees from Chongqing University. He is a professor and Ph.D. supervisor in Chongqing University. His main research interests include distributed computing, operating system and embedded system. (Email: sychen@cqu.edu.cn)

TIAN Yingming was born in 1979. He received the B.S. degree from Shenyang University of Technology, received the M.S. degree from Chongqing University of Post and Telecommunication. He is now a Ph.D. candidate in Chongqing University. His research interests include pattern recognition and artificial intelligence.

WU Peng was born in 1963. He received the B.S. degree from Hefei University of Technology, received the M.S. and Ph.D. degrees from Chongqing University. He is a professor and Ph.D. supervisor in Chongqing University. His research interests include pattern recognition and artificial intelligence. (Email: wupeng@cqcy.com)