Prediction of Sepsis from Clinical Data Using LSTM and XGBoost

Yongchao Wang[1], Bin Xiao[1*], Xiuli Bi[1], Weisheng Li[1], Junhui Zhang[2*], Xu Ma[3]

[1] Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing, China
[2] The First Affiliated Hospital of Chongqing Medical University, Chongqing, China
[3] Human Genetics Resource Center, National Research Institute for Family Planning, Beijing, China

Abstract

Sepsis is a life-threatening condition, and more than 6 million people die from sepsis every year. Therefore, developing an objective and efficient computer-aided tool for early detection of sepsis has become a promising research topic. In this paper, we present two methods for early prediction of sepsis from clinical data. One is neural network-based method and the other is XGBoost-based method. Considering the temporal relationship between clinical data from sepsis patients in the ICU, we built a Long Short-Term Memory (LSTM) network to extract the intrinsic relation between different indicators in clinical data and meanwhile model the temporal dependencies, which only uses the previous information not future information to predict the results. Since neural networks have made great achievements in unstructured data, such as image processing and speech processing, while traditional machine learning methods are better at processing structured data than neural networks. In addition, we trained an XGBoost model on the pre-processed data for improving the prediction accuracy. We only used the first seven vital signs in our network and in official phase, and the LSTM-based method has the utility score is 0.267 and the score of XGBoost-based method is 0.392.

1. Introduction

Since the concept of sepsis (i.e. Sepsis 1.0) [1] was proposed by the American College of Chest Physicians (ACCP) and the Society of Critical Care (SCCM) in 1991, research on sepsis has increased in recent years. In 2001, SCCM, ACCP, and the European Society for Critical Care Medicine (ESICM) held a joint meeting in Washington to revise Sepsis1.0, developed guidelines for medical treatment, and proposed a new diagnostic standard, Sepsis 2.0 [2]. But Sepsis 2.0 is too complicated, so it is rarely used clinically. In 2016, top scholars from the United States, Europe and Australia made a special group and proposed Sepsis 3.0 [3] based on big data analysis. Sepsis 3.0 is defined as a life-threatening organ failure caused by the body's uncontrolled response to infection. Sepsis 3.0 uses SOFA definition organ failure and propose the concept of Quick SOFA (qSOFA) for quickly and easily to assess risk of suspected infection or clinical deterioration.

In recent years, despite the significant advances in anti-infective treatment and organ function support technology, the mortality rate of sepsis is as high as 30%-70%. Moreover, the cost of treatment for sepsis is high and the medical resources are expensive. The annual medical expenses incurred by sepsis for U.S. hospitals are as high as $24 billion (13% of U.S. medical expenses), which exceeds any other health condition and most of the costs are for patients who have not been diagnosed with sepsis at the time of admission [4].

Each hour of delayed treatment has been associated with roughly a 4-8% increase in mortality [5]. Therefore, early diagnosis of sepsis is of great significance, not only can control the disease earlier, but also reduce unnecessary expenses for patients. In order to use the clinical data to early predict sepsis, we propose two methods, one is neural network-based method using LSTM algorithm, and the other is XGBoost [6] that is a traditional machine learning-based method.

2. Challenge Data analysis

The open training data published in challenge came from ICU patients in two independent hospitals [7]. The data for each patient will be contained within a single pipe-delimited text file. Each file has the same header and each row represents a single hour's worth of data. Available 40 features consist of Demographics, Vital Signs, and Laboratory values. Statistically speaking, there are a total of 40,336 patient measurement records in the training set, and each row of records contains a single hour's indicator data for the patient. The statistics show that there are 1,552,210 rows of data in the training set, and the imbalance of samples in the datasets is very serious. As shown in Figure 1, only 2932 of the 40336

subjects had sepsis, accounting for 7.27%, and the data collected from patients with sepsis was only 27,916 rows, just only 1.8% of all data.

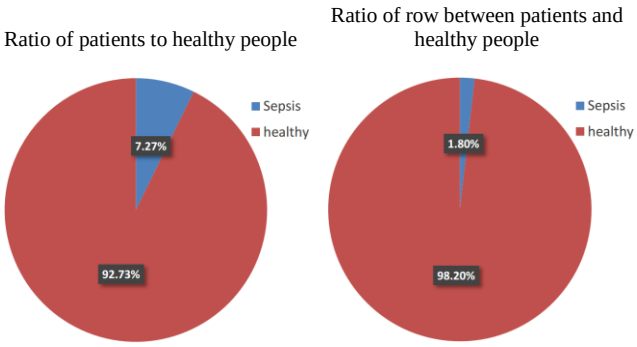

Figure 1: Ratio of patients to healthy people (left) and ratio of row between patients and healthy people (right).

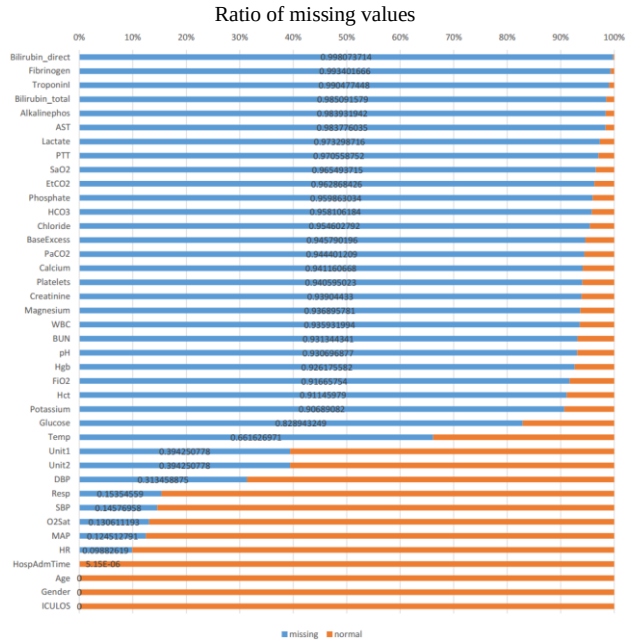

Figure 2: Ratio of missing values for each feature, the blue portion of the percentage stacked bar chart indicates the proportion of missing values.

Serious sample imbalance is one of characteristics of this dataset, and another characteristic is that the number of missing values is large.

As shown in Figure 2, 26 features have more than 90% of the values missed, even 6 features have more than 98% values missed. There are only 9 features with missing values less than 20%, and only 3 features (age, gender, ICULOS i.e. hours since ICU admit) are with complete data.

The third characteristic of this database is that each patient has a different recording time. The longest recording time is 336 hours and the minimum time is 8 hours.

In summary, the three characteristics of the database increase the difficulty of finally prediction of sepsis.

3. Data processing

Based on the characteristics of the database, we perform two data pre-processing strategies: missing value filling and feature expansion.

3.1. Missing value filling

Due to the large amount of data missing and the inability to use post-time data when predicting whether there is sepsis at current time, the filling strategy we performed was as follows:

1) Along the timeline, when a data missing was encountered, the currently missing data was filled with the last non-missing data of previous data.
2) In LSTM-based method, the missing data remaining after step 1 was filled with value 0. Since XGBoost allows the existence of missing values, this pre-processing step is not performed in XGBoost-based method.

3.2. Feature expansion

We implemented two feature expansion methods. Considering that the early diagnosis of sepsis is not only beneficial to the treatment of patients, but also reduces the economic burden of patients. The challenge encourages the prediction of sepsis with 6 hours early in advance.

The training data is physiological and biochemical measurement data for each hour. The changes in physiological and biochemical indicators can reflect the health of the human body. This is an important basement for judging whether the human body is suffering from sepsis. In order to extract changes in physiological and biochemical indicators, the difference before 6 hours was calculated from the seventh hour. And not all of the features given by the training data are suitable for calculating the difference. We carefully selected 34 features for the extension of the feature difference.

Sepsis 3.0 emphasizes that Sepsis is the host's uncontrolled response to infection and life-threatening organ dysfunction. According to the Sepsis 3.0 standard, sepsis satisfies the following formula:

$$sepsis = infection + (SOFA > 2) \quad (1)$$

In 1994, ESICM presented the sequential organ failure assessment (SOFA) score in Paris [8]. The SOFA score can dynamically assess the condition of organ failure using limited routine measurements. According to some physiological indicators, the functional scores of 6 organ systems were estimated which to range from 0 (no organ dysfunction) to 4 (severe organ dysfunction), and the

individual organ scores are then summed to a total score [9]. Combined with the features given by the training data, we designed 5 new features related to SOFA. The extended SOFA score features are calculated based on Table 1. In addition, we also calculated the sum of 4 SOFA scores.

Table 1: Conversion table for SOFA related features.

	0 score	1 score	2 scores	3 scores	4 scores
Platelets (count*10^3/μL)	>150	<=150	<=100	<=50	<=20
Bilirubin_tota (mg / dL)	<1.2	>=1.2	>=1.9	>=5.9	>=11.9
Creatinine (mg/dL)	<1.2	>=1.2	>=1.9	>=3.4	>=4.9
Mean arterial pressure (mm Hg)	>=70	<70	-	-	-

The qSOFA uses three of the most effective indicators for predicting poor prognosis in patients with Sepsis: respiratory rate (RR), Glasgow Coma Scale (GCS), and systolic blood pressure (SBP). Since the latest definition of septic shock focuses on lactate levels, we calculated three differences in this paper, and Table 2 shows the indicator and threshold.

Table 2: Selected indicators and corresponding thresholds

	selected indicator	threshold
SBP	(mm Hg)	100
RR	(beats per minute)	22
Lactate	(mmol/L)	2

In summary, we got 82 features, 40 are given by training data, and 42 are obtained by expansion.

4. Method

Our proposed scheme includes two methods that are LSTM-based and XGBoost-based. The former can automatically extract features, the latter rely on artificial features design. We attempted to compare these two method to find out which one is more suitable for predicting sepsis form clinical data.

4.1. LSTM

Because most of the features have the characteristic of very serious values missing, it is not easy for training the LSTM. Moreover, considering the difficulty of data collection, here, we just use seven vital signs to train the LSTM model.

LSTM is a neural network designed to process time series data, and the training data in the sepsis database includes hourly physiological and biochemical indicators. Therefore, considering the time dimension, it is very reasonable to use LSTM to treat sepsis data.

$$X^T = X^t \qquad t \in [1,\ T] \tag{2}$$

$$X^t = [x_1^t, x_2^t, ...x_i^t, ..., x_7^t] \qquad i \in [1,\ 7] \tag{3}$$

Where suppose X^T is the data of a patient, T is the number of rows in a pipe-delimited text file, and each person is different, $X^t \in \mathbb{R}^7$ is a vector represents a row of data, and i denotes the i-th feature for training the LSTM.

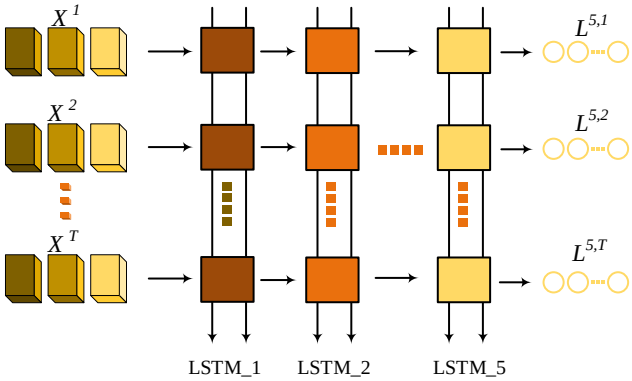

Figure 3: 5-layer LSTM network, in this figure, LSTM is a schematic diagram of the expansion along time.

As show in Figure 3, we built a 5-layer LSTM network and the number of hidden neurons is 14, 28, 56, 112, and 224, respectively. LSTM accepts a sequence and processes it through internal neurons to output another sequence. For example, suppose $X^1 = [x_1^1, x_2^1, ..., x_7^1]$ is the first row of data record by a patient. A new sequence $L^{1,1} \in \mathbb{R}^{14}$ is output after the first layer of LSTM (i.e. LSTM_1) processing.

$$L^{1,1} = \partial(H(X^1)) \tag{4}$$

Where $H()$ stands for processing of LSTM network, and $\partial()$ stands for relu activation function.

In a similar fashion, LSTM_5 output a vector $L^{5,1} \in \mathbb{R}^{224}$.

$$L^{5,1} = \partial(H(L^{4,1})) \tag{5}$$

When the LSTM processes the data after the first line, the recurrent connections allow a memory which comes from previous inputs to influence the final network output. In order to get the final prediction, a fully connected layer with two neural units was built after LSTM. And a weighted cross-entropy function was adopted, learning rate was set to 0.01.

Due to the sample imbalance is too serious, and in order to increase the probability that one sample was predicted to sepsis, the final prediction probability of disease is the network prediction probability plus a constant 0.00024079.

4.2. XGBoost

XGBoost is a boosting algorithm, and works very well for dealing with structured data classification problems. The idea of the boost algorithm is to integrate a series of weak classifiers into a strong classifier and XGBoost builds many CART regression trees based on sample features and integrates them for final prediction. Specifically speaking, XGBoost randomly selects features based on the pre-set parameters and automatically selects the appropriate threshold to build a CART tree. Then iteratively repeats the tree-building operation until the number of trees reaches maximum depth which pre-set. Meanwhile, the early stopping technique can also be used to stop training when the increase in the number of trees is no longer bringing gain. According to the structure of the CART tree and the sample eigenvalues, each tree assigns each sample to the leaf node, and XGBoost calculates the final prediction result based on the predicted values of the leaf nodes of all the CART trees.

We use all 82 features to train XGBoost, in our model, the number of trees in the XGBoost is 16, the maximum depth of the tree is 10, learning rate is 0.01, minimum leaf node weight is 1, subsample rate is 0.8, column sample by tree is 0.8, column sample by level is 0.9, "reg_alpha" is 1, "reg_lambda" is 0.5, and "scale_pos_weight" is 5.5. We empirically select the threshold as 0.24, and the sample was predicted to sepsis when the prediction probability of the model is greater than the threshold.

5. Result

In official phase, the method based on LSTM achieves utility score 0.267 and the method based on XGBoost is 0.392. In the local five-fold cross-validation, the former got a utility score of 0.314, and the latter obtained 0.343.

6. Conclusion

In this article, two methods that LSTM-based and XGBoost-based for predicting sepsis are proposed. By comparison, the second method has a higher utility score. We think the reason is that a large number of values in the training data are missing and a serious sample imbalance, which makes the neural network-based method very difficult to train. Since XGBoost is better at handling structured data and has a missing value processing mechanism, XGBoost achieves better results after feature filling and feature expansion.

Acknowledgments

This work was partly supported by the National Science & Technology Major Project (2016YFC1000307-3), the National Natural Science Foundation of China (61572092), the Natural Science Foundation of Chongqing (cstc2018jcyjAX0117, cstc2016jcyjA0407), the Scientific & Technological Key Research Program of Chongqing Municipal Education Commission (KJZD-K201800601), and the Chongqing research and innovation project of graduate students (CYS18245).

References

[1] American College of Chest Physicians, Society of Critical Care Medicine Consensus Conference Committee. American College of Chest Physicians/Society of Critical Care Medicine Consensus Conference: definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis [J]. Crit Care Med, 1992, 20: 864-874.

[2] Levy M M, Fink M P, Marshall J C, et al. 2001 sccm/esicm/accp/ats/sis international sepsis definitions conference [J]. Intensive care medicine, 2003, 29(4): 530-538.

[3] Singer M, Deutschman C S, Seymour C W, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3) [J]. Jama, 2016, 315(8): 801-810.

[4] Paoli C J, Reynolds M A, Sinha M, et al. Epidemiology and costs of sepsis in the United States—An analysis based on timing of diagnosis and severity Level[J]. Critical care medicine, 2018, 46(12): 1889.

[5] Kumar A, Roberts D, Wood K E, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock [J]. Critical care medicine, 2006, 34(6): 1589-1596.

[6] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, 2016: 785-794.

[7] Reyna MA, Josef C, Jeter R, Shashikumar SP, M. Brandon Westover MB, Nemati S, Clifford GD, Sharma A. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. Critical Care Medicine, in press.

[8] Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. Intensive Care Med. 1996; 22:707–10.

[9] de Grooth H J, Geenen I L, Girbes A R, et al. SOFA and mortality endpoints in randomized controlled trials: a systematic review and meta-regression analysis[J]. Critical care, 2017, 21(1): 38.

Address for correspondence:

Bin Xiao
School of Computer Science
No.2, Chongwen Road, Nan'an district, Chongqing, China
xiaobin@cqupt.edu.cn
Junhui Zhang
The First Affiliated Hospital of Chongqing Medical University
No. 1 Yixueyuan Road, Yuzhong District, Chongqing, China
2275610878@qq.com