Early Prediction of Sepsis Using Multi-Feature Fusion Based XGBoost Learning and Bayesian Optimization

Meicheng Yang[1], Xingyao Wang[1], Hongxiang Gao[1], Yuwen Li[1], Xing Liu[2], Jianqing Li[1*], Chengyu Liu[1*]

[1] State Key Laboratory of Bioelectronics, School of Instrument Science and Engineering, Southeast University, Nanjing, China
[2] Department of Anesthesiology, Third Xiangya Hospital, Central South University, Changsha, China

Abstract

Early prediction of sepsis is critical in clinical practice since each hour of delayed treatment has been associated with an increase in mortality due to irreversible organ damage. This study aimed to develop an algorithm for accurately predicting the onset of sepsis in the proceeding of six hours. Firstly, we selected 37 available variates features after data pre-processing, and then extracted three kinds of features as well in this paper, including 62 missing value features, 8 scoring quantified features and 61 time series features. After that, a multi-feature fusion based XGBoost classification model was developed and was further improved by a Bayesian optimizer and an ensemble learning framework. Analysis was performed on the PhysioNet/Computing in Cardiology Challenge 2019, which provided a publicly available sepsis data sourced from 40,336 ICU patients. Finally, after searching an optimized predicted risk threshold of 0.522 through the official submissions, our team "SailOcean" applied the developed model on the full hidden test set of 24,819 ICU patients from three hospital systems and obtained a final $U_{normalized}$ score (U-Score) defined by the organizers of 0.364, which was the highest unofficial score.

1. Introduction

Sepsis is one of the most common critical conditions in the emergency department, and occurs when the body loses control of its response to infection [1]. It has always been a major focus in clinical and basic research of critical care medicine, because of its severe morbidity, mortality and medical costs. Traditionally, rule-based disease severity scoring systems such as SOFA [2], qSOFA [1], NEWS [3], APACHE II [4] etc. have been proposed to define sepsis in hospitals, but they don't meet the urgent need for early sepsis detection to get effective treatment.

Nowadays, the increase in publicly available electronic health records (EHRs) [5] has brought tremendous opportunities in developing data-driven and efficient machine learning models to help diagnose diseases. Therefore, in recent years, to achieve the goal of early prediction of sepsis using physiological data, researchers have proposed many rule-based machine learning or deep learning models [6]. However, direct comparison of these methods is not possible due to the differences between clinical criteria, available patient variables, predictive tasks, evaluation metrics and so on [7].

The PhysioNet/Computing in Cardiology Challenge 2019 focuses on the early prediction of sepsis from multi-measurement clinical data [7]. In this paper, we develop a real-time algorithm using multi-feature fusion based XGBoost [8] learning and Bayesian optimization [9] to predict sepsis 6 hours before the clinical definition of sepsis. A total of 168 features are extracted from the available patient variates to train our XGBoost model and further improve the performance by using a Bayesian optimizer and an ensemble method. After verifying the effectiveness of the proposed algorithm and discussing the impact of predicted risk threshold on local test set formed by ourselves, we create the final ensemble model on the entire public challenge database via 5-fold cross validation and evaluate it on the hidden test set.

2. Methodology

The framework of our proposed algorithm about the early prediction of sepsis is shown as Figure 1. Raw patient's data is analyzed first for helping us to get more effective information in feature extraction. Then we divide the data set into train set, validation set and local test set. After that, XGBoost classifier receives training data after feature extraction as inputs, and tunes hyperparameter automatically using a Bayesian optimizer. Meanwhile, the 5-fold cross validation method is used to verify the stability of our algorithm and exports an ensemble model. Finally, we discuss the impact of predicted risk threshold on the $U_{normalized}$ score (U-Score). The details of specific algorithm implementation are as follows.

 ISSN: 2325-887X DOI: 10.22489/CinC.2019.020

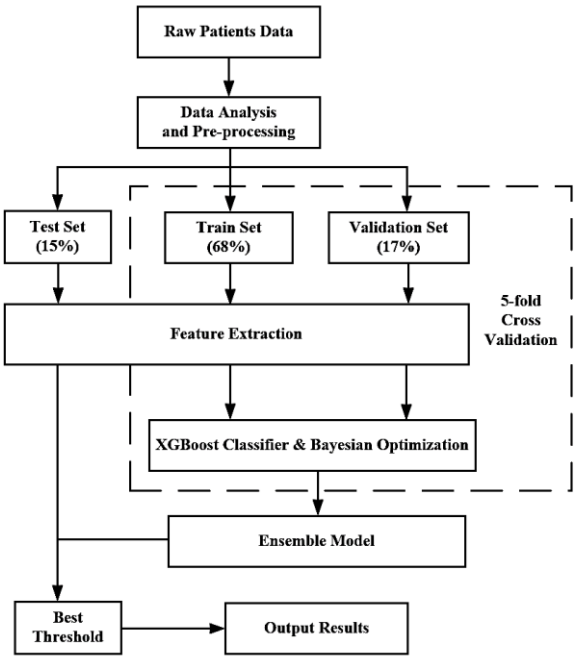

Figure 1. Framework of proposed algorithm

2.1. Data analysis and pre-processing

The public challenge database from two hospital systems included a total of 40,336 patients, 2,932 (7.27%) of whom developed sepsis with a longer mean recording time compared with non-septic patients (58.6 vs 38.9 hours). Among the available patient variates, missing value of each vital signs is less than 16% except Temperature (Temp) of 66%, Diastolic BP (DBP) of 31% and End tidal carbon dioxide ($EtCO_2$) of 96%. For laboratory values, except for Glucose missing value of 83%, the other measured variates are missing more than 90%. Both of the two administrative identifiers for ICU unit (Unit 1 and Unit 2) miss values nearly 39%, while others are fixed variables for the demographics.

After the data analysis, we pre-process the raw data in the following two steps.

(1) Step 1: Remove the variates such as Bilirubin_direct, Troponin I and Fibrinogen due to whose missing values account for more than 99%.

(2) Step 2: Impute missing data with forward-filling strategy. If there is one previous recorded value of variate *v* at time step $t_p < t$, we perform forward-filling by setting $x_v^{(t)} = x_v^{(t_p)}$ to handle the missing value of *v* at time step *t*. However, if there is no previous recorded measurement or the variable is missing entirely, this kind of null values will not be processed.

2.2. Dataset partition

The raw dataset of 40,336 patients is divided by septic and non-septic patients separately, the results of dataset partition are shown in Table 1. The local test set is formed with a fixed size by ourselves, while using a 5-fold cross validation method obtains the train and validation sets.

Table 1. Results of dataset partition

Dataset		Sepsis	Non-sepsis	Total
5-fold	Train set	1,994	25,435	27,429
	Validation set	498	6,358	6,856
Local test set		440	5,611	6,051

2.3. Feature extraction

Apart from the selected 37 variable features after data pre-processing, we extract three kinds of features as well in this section, including 62 missing value features, 8 scoring quantified features and 61 time series features.

2.3.1 Missing value features

The recording time of clinical variates varies with patients and even over time, so there are many missing values in physiological records as well as some completely missing variates. However, the data is not randomly missing especially in ICU because it may reflect the clinician's decision related to the severity of patients [10]. Meanwhile, in the public challenge database, we observe that the average proportion of missing values of each measurement variate in septic patients is lower than that in non-septic patients after Step 1 in data pre-processing. Thus, we design two missing data indicator sequences to excavate the potential predictive information of the missing data. This process is performed on 31 variables after the demographics were excluded from the selected 37 variable features after data pre-processing.

(1) Measurement frequency sequence: Record the number of variable measurements before the current time.

(2) Measurement time interval sequence: Record the time interval from the last measurement between the current time. We set -1 at the moment when there is no previous recorded measurement.

An example of the two missing data indicator sequences is shown in Table 2. The first row represents an eight hours' time series of Temp, the second row indicates the measurement frequency sequence while the last row indicates the measurement time interval sequence.

Table 2. Example of the missing data indicator sequences

nan	38.0	38.1	nan	nan	38.2	nan	37.4
0	1	2	2	2	3	3	4
-1	0	0	1	2	0	1	0

2.3.2 Scoring quantified features

The magnitude of the measurements reflects the response of the human physiological system to infection. We highlight the importance of several measurements to quantify abnormalities according to some scoring system. The qSOFA score is identified as 1 with Systolic BP (SBP) $\leq$ 100 mm Hg and Respiration rate (Resp) $\geq$ 22/min, otherwise 0. The measurements of Platelets, Bilirubin, Mean arterial pressure (MAP), and Creatinine are scored respectively under the rules of SOFA score, while Heart rate (HR) and Temp are scored on the basis of NEWS score.

2.3.3 Time series features

In order to obtain dynamic changes from patients' recording sequence, two kinds of time series features are calculated as below.

(1) Differential features: These features are extracted by calculating the difference between the current recorded value and the previous last measurement.

(2) Sliding window-based statistical features: The five measurements including HR, Pulse oximetry (O_2Sat), SBP, MAP and Resp are selected for sliding window processing because they have the least missing values. First, a fixed-length six-hour sliding window is applied to segment each record with a step of one hour as shown in Figure 2. For the window less than 6 hours in length, the last hour of data is filled repeatedly until the length of window is 6. For example, if gave only 4 hours' data, we fill the window time series into t = [0,1,2,3,3,3]. After that, to achieve the goal of early sepsis detection, if $t_{optimal}$ (6 hours before the onset time of sepsis) falls within the window, this segment is labelled as 1, otherwise 0. Finally, we calculate some typical statistical features including the maximum, minimum, mean, median, standard deviation and differential standard deviation of every measurement in each window.

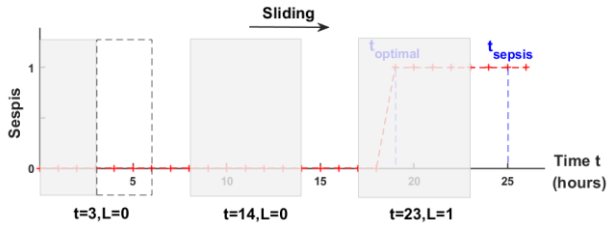

Figure 2. Process of sliding window

2.4. Classification

2.4.1 XGBoost and Bayesian optimization

XGBoost was proposed in 2015 and has been a widely used tool in data mining contests with great success [8]. It is a decision-tree-based ensemble machine learning algorithm that uses a gradient boosting framework. The algorithm supports parallelization in tree construction and is robust enough by using a more regularized model formalization. In addition, it has a specific processing method for sparse data which is important in our classification task with massive missing data.

Generally, hyper-parameter optimization aims at looking for the best hyper-parameter values to minimize the objective loss function. In our algorithm, Bayesian optimization with Tree Parzen Estimator is used to build a probability model and select the most promising hyperparameters. For XGBoost, max depth, learning rate, colsample_bytree and regularization coefficient are optimized in our model.

2.4.2 Model training

Using our feature extraction approach will result in over 1 million hours of data in the training process. However, only roughly 1.8% of these data corresponds to a positive outcome. Consequently, in order to deal with the serious class imbalance, a systematic way is provided by down sampling the excessive data instances of the majority class in each cross validation.

Five different XGBoost classifiers are obtained by 5-fold cross validation. Each classifier is learned to realize early prediction of sepsis via the training dataset and gets improvement using Bayesian optimization via the validation dataset. Specify the objective option as "binary:logistic" in XGBoost for binary classification and probability output. Furthermore, early-stopping strategy is also used to avoid over-fitting. After that, we ensemble the five XGBoost models by averaging their output probabilities to make the final decision.

2.4.3 Model evaluation

To evaluate the effectiveness and reliability of our proposed ensemble model for the early prediction of sepsis, we compute the U-Score of the five individual XGBoost models and the ensemble model on the local test set, respectively. Moreover, the predicted risk threshold may have an impact on rewarding or penalizing our algorithm, so we conduct an extra experiment to search an optimal predicted risk threshold to output results.

3. Results and discussions

To make a horizontal comparison between the five individual XGBoost models trained by 5-fold cross validation and the ensemble model, we set the predicted risk threshold as 0.50 for sepsis 0/1 classification firstly. Then we report the results of an area under receiver operating characteristic (AUROC), classification accuracy (ACC) and U-Score. Table 3 presents the performance of different models on the local test set, which is formed by

ourselves according to Table 1. There is no significant difference between the results which indicates the stability of our algorithm. In addition, we can observe that the ensemble model gains the best ACC, AUROC and U-Score at the same time. Therefore, the results confirm that using ensemble method allows to produce better predictions compared to a single model.

Table 3. Performance of different models on local test set formed by ourselves

Model		ACC	AUROC	U-Score
Individual XGBoost Model	1	0.807	0.838	0.409
	2	0.814	0.838	0.408
	3	0.818	0.844	0.411
	4	0.810	0.842	0.406
	5	0.813	0.839	0.419
	Average	0.812	0.840	0.411
Ensemble Model		0.818	0.847	0.425

Afterwards, we set different predicted risk thresholds range from 0.49 to 0.55 for the ensemble model and get the results as shown in Figure 3. The figure demonstrates that the threshold has a certain impact on the result of U-Score.

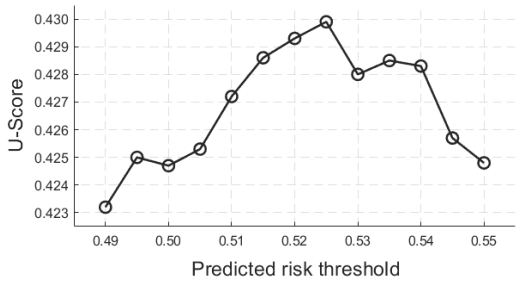

Figure 3. Performance of the ensemble model with different predicted risk thresholds when applied on the local test set formed by ourselves

Finally, our submitted training ensemble model is created on the entire 40,336 patients via 5-fold cross validation, and we determine the optimal predicted risk threshold as 0.522 after testing our model on the subset of the hidden test set during official phase. The performance of our method when applied on the hidden test set is shown in Table 4. The final results show that the method of our team "SailOcean" yields the overall U-Score of 0.364 when tested on the full hidden test set of 24,819 patients from three hospital systems.

Table 4. Final performance on hidden test set

Hidden test set	A	B	C	Full
U-Score	0.430	0.422	-0.048	0.364

4. Conclusion

In this paper, to address the issue of sepsis early prediction, our team "SailOcean" proposed a feasible and open-source algorithm using multi-feature fusion based XGBoost learning and Bayesian optimization. This algorithm can predict the risk of sepsis at each hour in time using only data until that moment, which is vitally important in practical clinical use for life-saving. Results show that when applied on the full hidden test set, our method obtains an overall U-Score of 0.364, which is the highest unofficial score.

References

[1] Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche JD, Coopersmith CM, et al. The third international consensus definitions for sepsis and septic shock(Sepsis-3). Journal of the American Medical Association 2016;315(8):801-810.

[2] Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, Reinhart CK, Suter PM, Thijs LG. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive Care Medicine 1996;22(7):707-710.

[3] Smith GB, Prytherch DR, Meredith P, Schmidt PE, Featherstone PI. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. Resuscitation 2013;84(4):465-470.

[4] Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Critical Care Medicine 1985;13(10):818-829.

[5] Birkhead GS, Klompas M, Shah NR. Uses of electronic health records for public health surveillance to advance public health. Annual review of public health 2015;36:345–359.

[6] Calvert JS, Price DA, Chettipally UK, Barton CW, Feldman MD, Hoffman JL, Jay M, Das R. A computational approach to early sepsis detection. Computers in biology and medicine 2016;74:69-73.

[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 2019;In press.

[8] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In KDD 2016;785-794.

[9] Shahriari B, Swersky K, Wang Z, Adams RP, Freitas ND. Taking the human out of the loop: A review of Bayesian optimization. Proceedings of the IEEE 2016;104(1):148-175.

[10] Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. British Medical Journal 2009;338:b2393.

Address for correspondence:

Jianqing Li and Chengyu Liu
Sipailou Road 2, Southeast University, Nanjing, China
Email: ljq@seu.edu.cn and chengyu@seu.edu.cn