Sepsis Prediction in Intensive Care Unit Using Ensemble of XGboost Models

Morteza Zabihi[1], Serkan Kiranyaz[2], Moncef Gabbouj[1]

[1]Tampere University, Tampere, Finland
[2]Qatar University, Doha, Qatar

Abstract

Sepsis is caused by the dysregulated host response to infection and potentially is the main cause of 6 million death annually. It is a highly dynamic syndrome and therefore the early prediction of sepsis plays a key role in reducing its high associated mortality. However, this is a challenging task because there is no specific and accurate test or scoring system to perform early prediction. In this paper, we present a systematic approach for sepsis prediction. We also propose a new set of features to model the missingness in clinical data. The pipeline of the proposed method comprises three major components: feature extraction, feature selection, and classification. In total, 407 features are extracted from the clinical data. Then, five different sets of features are selected using a wrapper feature selection algorithm based on XGboost. The selected features are extracted from both valid and missing clinical data. Afterwards, an ensemble model consists of five XGboost models is used for sepsis prediction. The proposed algorithm is ranked officially as third place in the PhysioNet/Computing in Cardiology Challenge 2019 with an overall utility score of 0.339 on the unseen test dataset (our team name: Separatrix).

1. Introduction

Sepsis is defined as life-threatening organ dysfunction caused by a dysregulated host response to infection [1] and is often associated with lung, urinary tract, skin, and gut infections. The recent report of Center for Disease Control (CDC) shows that sepsis causes one out of every three hospital deaths [2] [3]. Besides the high mortality rate of sepsis, it imposes immense challenges to healthcare systems. From an economic perspective, sepsis implies high costs of hospital care with almost 17 billion USD annually in the United States [4] and 2.5 billion pounds in the UK [5]. Thus, early prediction of sepsis is a crucial element for appropriate clinical management and improvement of clinical outcomes.

The recent clinical criteria of sepsis [1] in the general hospital ward setting, recommend that quick Sequential (Sepsis-related) Organ Failure Assessment (qSOFA) should be used as a rapid evaluation of sepsis risk. This means that the patient should have at least two of the following clinical criteria to be considered as a patient with suspected infection: respiratory rate of 22 per minute or greater, altered mentation, and systolic blood pressure of 10 mmHg or less. Moreover, in [1], the SOFA$\geq$ 2 score is determined to represent organ dysfunction. SOFA score monitors laboratory values and vital signs such as the fraction of inspired oxygen (FiO2), the partial pressure of oxygen (PaO2), platelets, liver bilirubin, and mean arterial pressure [1]. However, sepsis is a dynamic condition, and such criteria may not meet or present in all the time. This leads to inaccurate results of such approaches [6]. In addition, using clinical criteria for sepsis diagnosis in patients with critical situations (e.g., ICU patients) can be even more challenging due to the misleading symptoms caused by other diseases [7].

Despite the slow changes in sepsis definitions, several studies have focused on the development of Machine Learning models to overcome the aforementioned challenges. In [8], the proposed method achieved significantly higher accuracy compared to the three standard sepsis-related scoring systems (i.e., SOFA, qSOFA, and MEWS). In [9], a variant recurrent neural network model is proposed for sepsis prediction. Their proposed model revealed that ICU length-of-stay, heart rate, white blood cell count, and temperature are the most relevant features for sepsis prediction. In [10], a model based on Weibull-Cox proportional hazards mode is proposed to predict the onset of sepsis in an ICU patient 4 to 12 hours prior to clinical recognition. Their method achieved the area under the receiver operating characteristic (AUROC) between 0.83–0.85. These models have achieved higher accuracy compared to traditional clinical criteria. However, further studies are needed to improve the robustness, false alarm rate, and interpretability of such models.

In this paper, we explore the use of an ensemble learning technique (Figure 1) for sepsis prediction in ICU. The main contributions of this study are:

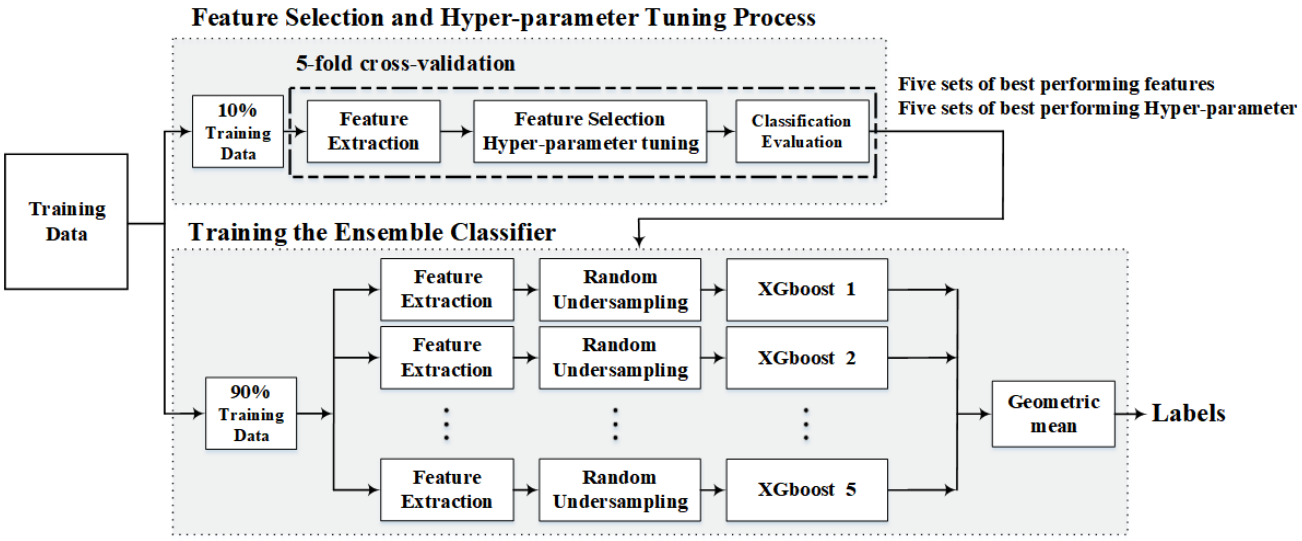

Figure 1. The training strategy of the proposed method

1) Investigating a comprehensive set of features and tracking the top clinically relevant features.
2) Introducing discriminative features for revealing the patterns of missing values in clinical data.
3) Designing a predictive model by ensembling 5 classifiers.

The remainder of this paper is organized as follows: In Section 2, the dataset is briefly described and the proposed method is explained. In Section 3, the evaluation results are presented and discussed. Finally, concluding remarks are outlined in Section 4.

2. Materials and methods

The dataset used in this competition is collected from 63097 ICU patients in three distinct hospitals. The training set includes 40336 records from two hospitals (hospitals *A* and *B*), while the remaining 22761 patient records (from hospitals *A*, *B*, and *C*) are kept hidden to be used for final ranking. For each patient, eight vital signs, six demographics variables, and 26 laboratory values are provided for every hour. More detailed information can be found in [11]. The feature extraction, feature selection, and classification approach are described next.

2.1. Feature engineering

Often the clinical data are not collected consistently. Therefore, it is expected that the majority amount of data for some covariates is missing. It has been shown that the imputation of missing values for such covariates does not significantly improve the prediction performance [12]. On the other hand, the missingness may convey useful information [13]. Therefore, in this work, two different types of features are extracted. The first type of feature targets the covariates with less than 70% of missingness while the second type of feature focuses on the patterns of missing values in clinical data. The combination of these features forms a set of 407 features in total (see Table 1). Once the features are extracted, they are normalized to a mean of 0 and unit standard deviation. The extracted features are described as follows:

The first type of features are extracted from 13 covariates of heart rate (HR), pulse oximetry (O2Sat), temperature (Temp), systolic blood pressure (SBP), mean arterial pressure (MAP), diastolic blood pressure (DBP), respiration rate (Resp), age, gender, administrative identifier for MICU unit (Unit1), administrative identifier for SICU unit (Unit2), hours between hospital admit and ICU admit (HospAdmTime), and ICU length-of-stay (ICULOS). Before extracting the first type of features, the missing value imputation is carried out by linear interpolation. If the data is less than 3 hours (i.e., less than three observations), then the missing values are replaced with the mean value of the corresponding covariate in the training data. For age and gender, the missing values are replaced by the first valid value. If all the values in the given observations were missing, then they are replaced by mean values. Once the imputation is performed, the following features are extracted:

(1) Sliding-window based features: Mean, minimum, maximum, median, variance, 95%, 99%, 5%, and 1% quantiles are calculated from the last 5 and 11 hours observations. We use two different time windows (i.e., 5 and 11 hours) to capture the short- and long-term temporal evolution of covariates.

(2) Non sliding-window based features: Energy, Shannon entropy, mean of the first differences, and the lengths of observations are calculated from the given observations.

(3) The last observation values of the 13 covariates are also used as 13 separate features.

Table 1. List of the extracted features

Type	Features	#features	
1 *Imputation, 13 covariates*	*Mean*, *minimum*, *maximum*, *median*, *variance*, *95%*, *99%*, *5%*, and *1% quantiles* from the last 5 and 11 hours.	198	245
	Energy, *Shannon entropy*, *mean of the first differences*, and *the lengths of observations*	34	
	last observation values of the 13 covariates	13	
2 *No imputation, 38 covariates*	*Mean* and *variance* of L_C	76	162
	Summation and *variance* of L_{CV}	76	
	Mean and *variance* of L_O	10	

To calculate the second type of features, age and gender are excluded from the given covariates. These two demographic variables are constant for each patient during the monitoring and therefore their absence does not convey any information. To represent the missingness, we define the *sequence* abstraction. Each *sequence* is defined as a set of consecutive measurements where the values are only either missing or present. Therefore, each *sequence* can only have missing or present values. For instance, let's imagine the SBP measurements for 6 hours are $\{nan, 122, 98, nan, nan, 123\}$, then based on the definition, we have 4 *sequences* of $\{nan\}$, $\{122, 98\}$, $\{nan, nan\}$, and $\{123\}$. Using the *sequence* abstraction the following features are calculated (see Figure 2):

(1) Mean and variance of the lengths of *sequences* along each covariate, L_C.

(2) Summation and variance of the lengths of *sequences* with only valid values (without missing) along each covariate, L_{CV}.

(3) Mean and variance of the lengths of *sequences* along each observation, L_O, in the last 5 hours.

It is worth mentioning that the input clinical data have varying lengths, and it is possible that the number of observations is not enough to extract the sliding-window based features. For such cases, the clinical data is padded using the first observation. The amount of padding equals to the difference between the number of observations in the given data and the number of needed ones. This enables us to transform the raw data into a feature space with a fixed length. Thus, discriminative methods, such as XGboost and random forest, can be applied to such dynamic data.

Observation	HR	Temp	SBP	L_O
1	NaN	NAN	NaN	{3}
2	97	NaN	98	{1, 1, 1}
3	89	NaN	122	{1, 1, 1}
	90	NaN	NaN	{1, 2}
	103	NaN	122	{1, 1, 1}
.	110	NaN	NaN	{1, 2}
.	108	36.11	123	{3}
.	106	NaN	93	{1 1, 1}
	104	NaN	133	{1, 1, 1}
10	102	NaN	134	{1, 1, 1}
L_C	{1, 9}	{6, 1, 3}	{1, 2, 1, 1, 1, 4}	
L_{CV}	{9}	{1}	{2, 1, 4}	

Figure 2. Sequence abstraction for HR, Temp and SBP covariates

2.2. Feature selection and classification

The proposed classification algorithm consists of two main steps:

(1) In the first step, five sets of best performing features and hyper-parameters are selected. We perform the feature selection and hyper-parameter tuning in a 5-fold cross-validation scheme using 10% of the original training data. For feature selection, we employ a wrapper feature selection algorithm based on XGboost (BoostARoota [14]). The importance metric is the number of times that a particular feature was split on in the XGboost algorithm. In addition, a grid search is used to find the best performing combinations of hyper-parameters.

(2) In the second step, we used an ensemble of five XGboost models. XGboost is a decision tree based ensemble using a gradient boosting framework [15] and its effectiveness has been established in a wide range of applications especially in prediction problems. To train the proposed ensemble, we randomly split the remaining 90% of the original data into five equally disjoint sets. Then, each set is used to train a distinct classifier. Moreover, due to the imbalance problem between sepsis and non-sepsis observations, we separately balance the data for each XGboost using the random undersampling technique. Finally, we use the geometric mean to integrate the outputs of the five classifiers. The training strategy of the proposed method is shown in Figure 1.

3. Results and discussion

We test our predictive model in a 5-fold cross-validation scheme using the training data. The results are reported in Table 2 (for more information about the score and metrics refer to [11]). The obtained utility scores (AUROC, AUPRC, F-measure) on the unseen test set *A*, *B*, and *C* are 0.422 (0.814, 0.102, 0.128), 0.395 (0.844, 0.110, 0.130), and -0.146 (0.793, 0.058, 0.044), respectively. Clearly, the performance of the proposed model on hospitals *A* and *B* (which are present in the training set) are robust with respect to our cross-validation. However, the performance drops drastically on the test set *C*. We believe that the main reason is that the missingness in hospital *C* has a different

Table 2. The results of the proposed method on the training data in a 5-fold cross-validation scheme and on the hidden test set. AUROC and ACC are area under the receiver operating characteristic and accuracy, respectively.

Fold	AUROC	ACC	Score
0	0.8387	0.8394	0.4366
1	0.8357	0.8418	0.4412
2	0.8436	0.8477	0.4521
3	0.8221	0.8451	0.3899
4	0.8268	0.8464	0.4208
Average (std)	**0.8333 (0.0078)**	**0.8440 (0.0030)**	**0.4281 (0.0215)**
The hidden test data			**0.339**

pattern compared to other hospitals. Here, missingness represents human behavior in recording the covariates and does not convey medical information. Therefore, missingness should be used with caution. That said, it should be noted that all the contestants fail to achieve a high score on test set C even if they have not used missingness information in their proposed methods.

Additionally, we observe that among the second type features (missingness) 102 out of 162 features were selected commonly using the BoostARoota algorithm. This shows the significance of the proposed features in sepsis prediction. Moreover, among the selected features, the HospAdmTime, the summation of L_{CV} for TMP, age, Unit1, variance of HR and TMP in the last 11 hours were ranked among the top 10 features.

4. Conclusions

In this work, we proposed a systematic approach for sepsis prediction in ICU. We investigate a set of features to capture the transitional states of covariates by using two time windows with different lengths. In addition, we introduce a new set of features to represent the missingness in clinical data. We examined the importance of features using the BoostARoota algorithm and found that the missing data convey relevant information for sepsis prediction in two out of three hospitals. The proposed method is officially ranked as the third team with a utility score of 0.339 on the unseen data (our team name: *Separatrix*).

References

[1] M. Singer, C. S. Deutschman, C. W. Seymour, et al., "The third international consensus definitions for sepsis and septic shock (Sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801-810, 2016.

[2] "Making health care safer: Think sepsis," *Center for Disease Control and Prevention (CDC)*, 2016.

[3] J. Hajj, N. Blaine, J. Salavaci, D. Jacoby, "The centrality of sepsis: A review on incidence, mortality, and cost of care," *Healthcare (Basel)*, vol. 6, no. 3, 2018.

[4] DC. Angus, WT. Linde-Zwirble, J. Lidicker, G. Clermont, J. Carcillo, MR Pinsky, "Epidemiology of severe sepsis in the United States: Analysis of incidence, outcome, and associated costs of care," *Crit Care Med*, vol. 29, no. 7, pp. 1303-10, 2001.

[5] M. Sanderson, M. Chikhani, E. Blyth, S. Wood, L. K. Moppett, T. McKeever, M. JR Simmonds, "Predicting 30-day mortality in patients with sepsis: An exploratory analysis of process of care and patient characteristics," *Journal of the Intensive Care Society*, vol. 19, no. 4, p. 299–304, 2018.

[6] Z. Zhang, NJ. Smischney, H. Zhang, S. Van Poucke, et. al., "AME evidence series 001-The society for translational medicine: Clinical practice guidelines for diagnosis and early identification of sepsis in the hospital," *Journal of Thoracic Disease*, vol. 8, no. 9, pp. 2654-2665, 2016.

[7] J. L. Vincent, "The clinical challenge of sepsis identification and monitoring," *PLoS Medicine*, vol. 13, no. 5, 2016.

[8] A. Mitra, K. Ashraf, "Sepsis prediction and vital signs ranking in intensive care unit patients," *arXiv*, 2018.

[9] S. P. Shashikumar, C. Josef, A. Sharma, S. Nemati, "DeepAISE--An end-to-end development and deployment of a recurrent neural survival model for early prediction of sepsis," *arxiv*, 2019.

[10] S. Nemati, A. Holder, F. Razmi, M. Stanley, G. Clifford, T. Buchman, "An interpretable machine learning model for accurate prediction of sepsis in the ICU," *Critical Care Medicine*, vol. 46, no. 4, pp. 547-553, 2018.

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

[12] N. Razavian, D. Sontag, "Temporal convolutional neural networks for diagnosis from lab tests," *arXiv*, 2015.

[13] J. H. Lin, P. J. Haug, "Exploiting missing clinical data in Bayesian network modeling for predicting medical problems," *Journal of Biomedical Informatics*, vol. 41, no. 1, pp. 1-14, 2018.

[14] C. DeHan, "BoostARoota," https://github.com/chasedehan/BoostARoota, 2017.

[15] T. Chen, C. Guestrin, "XGBoost: A scalable tree boosting system," in *International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA, 2016.

Address for correspondence:

Morteza Zabihi
P.O. Box 553, FI-33014, Tampere, Finland
Morteza.zabihi@tuni.fi