An Ensemble of Bagged Decision Trees for Early Prediction of Sepsis

Reza Firoozabadi, Saeed Babaeizadeh

Advanced Algorithm Research Center, Philips Healthcare, Andover, MA, USA

Abstract

Sepsis is a serious medical condition caused by the body's response to an infection. Early prediction and treatment of sepsis are critical. In response to the PhysioNet/CinC Challenge 2019, we developed an algorithm for early prediction of sepsis. Three datasets provided by the challenge are from ICU patients in three separate hospitals, two of which are publicly available to the participants, but the third is hidden and used for scoring. Data are highly unbalanced and contain many missing values. Each patient's data comprises hourly collected samples of 40 features. We preprocessed the data by a plausibility filter eliminating the outliers, forward filling of the missing data and replacing the remaining by population mean, and standardization of the numerical data. We developed an ensemble of bagged decision trees with a highly unbalanced misclassification cost to predict the sepsis for each sample of features in a patient. The classifier was trained on the first hospital dataset and validated on the second hospital dataset. A total of 15 important features and a set of hyperparameters were selected in an iterative training approach. Nine entries were submitted for evaluation of the utility score on a subset of the hidden dataset. The entry with the best utility score (0.335) was selected for running on the full test dataset and the final utility score was ??.

1. Introduction

According to the Sepsis-3 guidelines [1], "sepsis is a life-threatening organ dysfunction caused by a dysregulated host response to infection". This organ dysfunction is represented by two-point or more increase in the Organ Failure Assessment (SOFA) score. It is also discovered by records of clinical suspicion of infection in hospital either by ordering blood cultures or IV antibiotics.

Early prediction and treatment of sepsis are critical for reducing the mortality and morbidity, as well as the healthcare costs. Only in United States, more than 1.5 million cases of sepsis occur per year [2]. The significance of early prediction of sepsis cases and its impact on the survival rate of the patients has been presented in several papers [3-5]. Prompt identification of sepsis is recommended by clinical practice guidelines [6,7] and supported by studies suggesting that early treatment of sepsis reduces the mortality rate [3,8].

Although clinicians have proposed new definitions for sepsis, early detection and treatment of sepsis is still an issue and the limits of early detection are unknown. In order to address these issues, the organizers of the PhysioNet/Computing in Cardiology Challenge 2019 [9] set up a competition to develop automated open-source algorithms for the early detection of sepsis from clinical data. Three datasets were provided by the challenge from ICU patients in three separate hospitals, two of which are publicly available to the participants, while the third is hidden and used for scoring. The complications are the highly unbalanced data (1.8% sepsis onset) and high number of missing values (up to 99.8%) in public databases. The algorithm was evaluated according to its binary classification performance using a utility function created by the challenge organizers, rewarding for early predictions of sepsis and penalizing for too early/late/missing prediction of sepsis or false prediction of sepsis.

In response to this challenge, we developed an algorithm for early prediction of sepsis. We preprocessed the data by a plausibility filter, imputing, and standardization of the numerical data. A classifier modeled by an ensemble of bagged decision trees with a highly unbalanced misclassification cost function was developed to predict the sepsis for each sample of patient's features in time. The classifier was trained on the first hospital dataset and validated on the second hospital dataset. Important features and hyperparameters were selected in an iterative training approach.

The rest of this paper is organized as follows. In Section 2, we describe the method and material including the algorithm overview, database, data preprocessing, classifier, and feature selection. Section 3 provides the results. Discussion and conclusions are presented in section 4.

2. Method and Material

2.1. Algorithm Overview

Figure 1 shows the block diagram of the algorithm.

Patient's multi-feature data consist of several samples of features typically collected every hour. Data were preprocessed in several steps consisting of the outlier elimination, combination of the dependent variables, missing values imputation, and standardization.

The preprocessed data is then split into training and validation datasets. A tree-bagger classifier is trained using the training dataset and the important features are selected in an iterative approach until the best performance is achieved. The classifier is validated by validation dataset and the model is submitted for evaluation by a subset of the hidden test dataset. The classifier with the best utility score is selected for evaluation of the utility score on the full test dataset.

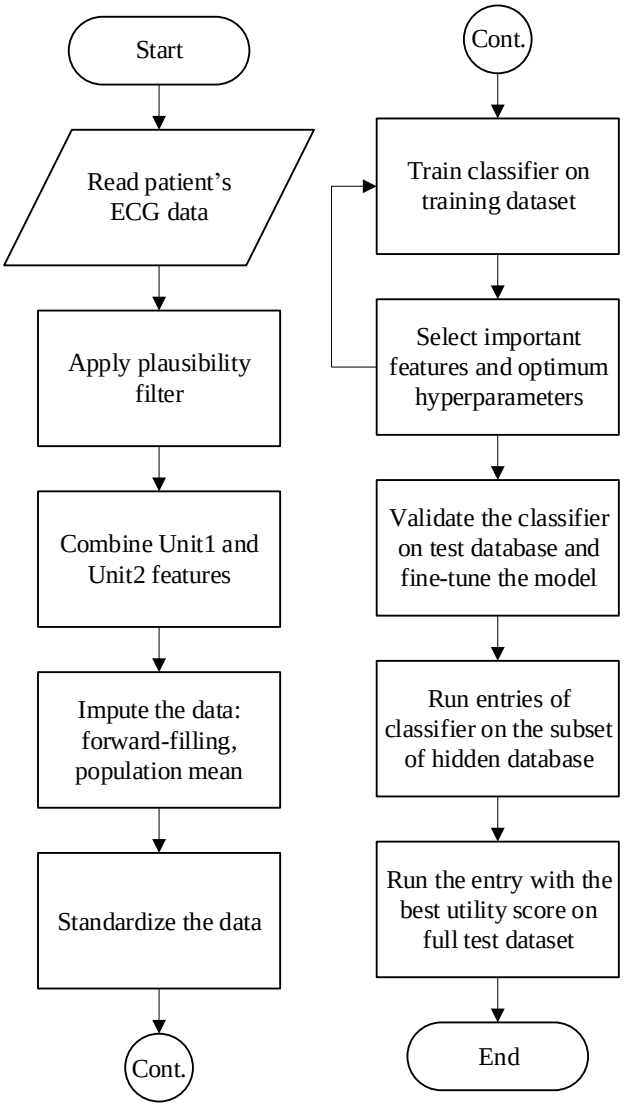

Figure 1. Block diagram of the algorithm consisting of data preprocessing (left branch) and evaluation procedure (right branch).

2.2. Database

The challenge provided three datasets from ICU patients in three separate hospitals. Datasets from two hospitals (A and B) were publicly available and used in the algorithm development. Datasets A and B contain 20,336 and 20,000 files, respectively. The third dataset was hidden and used for scoring and evaluation of the algorithm.

Each file in the datasets contains the records of one patient during the stay in ICU where samples were collected hourly, generating several time-series features. These features consist of three groups of vital signs (n=8), laboratory values (n=26), and demographic features (n=6). Some features show a high level of missing values.

Binary categorical features are gender, Unit1, and Unit2. Other features are numerical. Table 1 shows a list of features in each group of features. Statistic parameters including min, max, and mean, as well as the percentage of missing values in each feature are presented for all data from both public datasets A and B (n = 40,336) before preprocessing.

2.2.1. Data Preprocessing

Data were pre-processed before being used in development of the classifier model. The first step was applying a plausibility filter to the data. A range of valid values for each feature were defined based on its actual distribution and the knowledge in literature. Any value outside this range was assumed an outlier and marked missing for imputation. Table 1 presents the low and high values of the plausibility range for each feature.

Categorical features Unit1 and Unit2 are the administrative identifiers for ICU and are mutually exclusive; patient was either in MICU or SICU. In case of missing values, MICU (Unit1) was assumed.

There are lots of missing values in the features, ranging from 0% in some demographic features including age, gender, HospAdmTime, and ICULOS, to 99.8% in Bilirubin_direct. Percentage of missing values for each feature is shown in Table 1. Missing values were imputed by forward filling if the value was available in past. The remaining missing values with no past values were replaced by the population mean, calculated from the public datasets A and B after applying the plausibility filter. Numerical values were then standardized by reduction of the median values, divided by the standard deviation.

Figure 2 shows an example of data from a patient after preprocessing all features and combining Unit1 and Unit2 features. The preprocessed features were the imputed and standardized time series samples. In this example there are 39 features varying in a 54-hour interval.

Table 1. Statistical parameters for features in public datasets A and B. Missing percentage and the plausibility limits are also shown.

		Data in hospitals A and B				Plausibility	
	Features	**Min**	**Max**	**Mean**	**Missing (%)**	**Low**	**High**
Vital signs	**HR**	*20*	*280*	*84.6*	*9.9*	10	300
	O2Sat	*20*	*100*	*97.2*	*13.1*	60	100
	Temp	*20.9*	*50*	*37*	*66.2*	32	42.2
	SBP	*20*	*300*	*123.8*	*14.6*	40	280
	MAP	*20*	*300*	*82.4*	*12.5*	0	300
	DBP	*20*	*300*	*63.8*	*31.3*	20	130
	Resp	*1*	*100*	*18.7*	*15.4*	5	60
	EtCO2	*10*	*100*	*33*	*96.3*	0	150
Laboratory values	**BaseExcess**	*-32*	*100*	*-0.7*	*94.6*	-20	20
	HCO3	*0*	*55*	*24.1*	*95.8*	0	50
	FiO2	*-50*	*4000*	*0.6*	*91.7*	0	1
	pH	*6.62*	*7.93*	*7.4*	*93.1*	6	8
	PaCO2	*10*	*100*	*41*	*94.4*	0	200
	SaO2	*23*	*100*	*92.7*	*96.5*	0	100
	AST	*3*	*9961*	*260.2*	*98.4*	0	400
	BUN	*1*	*268*	*23.9*	*93.1*	0	500
	Alkalinephos	*7*	*3833*	*102.5*	*98.4*	0	250
	Calcium	*1*	*27.9*	*7.6*	*94.1*	0	20
	Chloride	*26*	*145*	*105.8*	*95.5*	75	145
	Creatinine	*0.1*	*46.6*	*1.5*	*93.9*	0	10
	Bilirubin_direct	*0.01*	*37.5*	*1.8*	*99.8*	0	50
	Glucose	*10*	*988*	*136.9*	*82.9*	0	1000
	Lactate	*0.2*	*31*	*2.6*	*97.3*	0	100
	Magnesium	*0.2*	*9.8*	*2.1*	*93.7*	0	10
	Phosphate	*0.2*	*18.8*	*3.5*	*96.0*	0	12
	Potassium	*1*	*27.5*	*4.1*	*90.7*	1	10
	Bilirubin_total	*0.1*	*49.6*	*2.1*	*98.5*	0	50
	TroponinI	*0.01*	*440*	*8.3*	*99.0*	0	200
	Hct	*5.5*	*71.7*	*30.8*	*91.1*	10	70
	Hgb	*2.2*	*32*	*10.4*	*92.6*	2	22
	PTT	*12.5*	*250*	*41.2*	*97.1*	0	250
	WBC	*0.1*	*440*	*11.4*	*93.6*	0	50
	Fibrinogen	*34*	*1760*	*287.4*	*99.3*	0	800
	Platelets	*1*	*2322*	*196*	*94.1*	5	1500
Demographics	**Age**	*14*	*100*	*62*	*0.0*	0	150
	Gender	*0*	*1*	*0.6*	*0.0*	0	1
	Unit1	*0*	*1*	*0.5*	*39.4*	0	1
	Unit2	*0*	*1*	*0.5*	*39.4*	0	1
	HospAdmTime	*-5367*	*23.99*	*-56.1*	*0.0*	none	none
	ICULOS	*1*	*336*	*27*	*0.0*	1	none

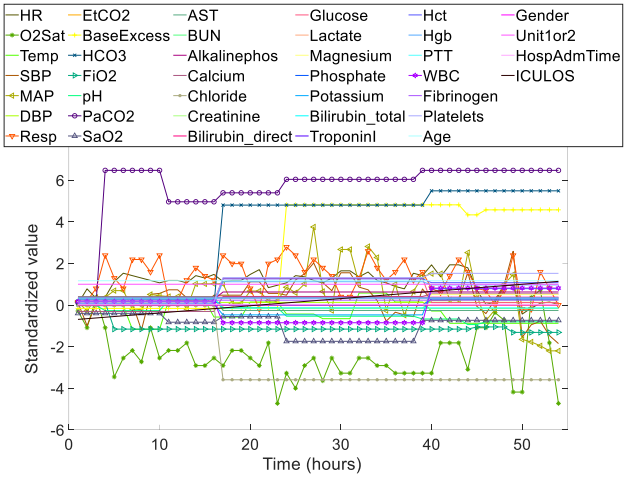

Figure 2. An example of a patient's preprocessed data with 39 features varying in the time.

2.3. Classifier

An ensemble of bagged decision trees was developed as the classifier with binary outputs: sepsis or no-sepsis. For each sample in time, the model accepts the important features as input. The ensemble consists of 100 decision trees. Due to the highly unbalanced nature of the data, a cost ratio of 1:37 was defined for misclassification of no-sepsis versus sepsis. Maximum number of splits is set to 100 with minimum leaf size of 3.

Figure 3 displays an example of a decision tree in the top panel and an ensemble of 100 trees in the bottom panel.

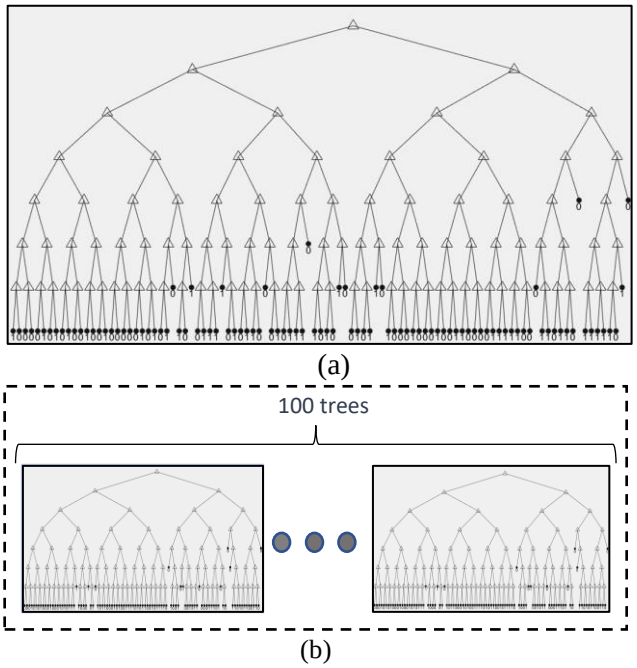

Figure 3. (a) Example of one decision tree. (b) Example of an ensemble of 100 decision trees.

2.3.1. Feature Selection

Important features were selected in an iterative approach finding the maximum utility score along with the optimized hyperparameters. A total of 15 features were selected which are listed in Table 2.

Table 2. Selected important features

Selected Important Features
HR
Temp
MAP
Resp
BaseExcess
FiO2
BUN
Calcium
Creatinine
Hct
WBC
Platelets
Unit1or2
HospAdmTime
ICULOS

3. Results

Our predictive model was trained using the features selected from dataset A in an iterative approach to find the best set of features and hyperparameters which result in the optimum performance. Each sample of the multi-feature data in each patient was used as the single input to the classifier resulting in a single output. Up to three initial samples in each patient with abundant missing values were discarded from analysis. Dataset B was used to validate and fine-tune the optimized model and the selected set of important features. A 10-fold cross-validation analysis on the training dataset (A) resulted in the average score utility 0.344888. A total of nine entries were submitted and the entry with the highest utility score (0.335174) was selected for running on the full hidden dataset. The algorithm was submitted to the competition organizers for running on a subset of hidden data to evaluate the utility score. The final utility score on the full hidden dataset was ??.

4. Discussion and Conclusions

We developed an algorithm for early prediction of sepsis in ICU patients up to 12 hours in advance using samples of features including the vital signs, lab results and demographics. The results show that the algorithm is successful in early prediction. Our algorithm uses single samples of the data. A sequence model such as a recurrent neural network (RNN) takes a series of samples as input and may improve the prediction performance. As future work, we are considering to develop an RNN such as an LSTM network and compare its performance to our model's.

References

[1] Singer M, Deutschman CS. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016 Feb 23;315(8):801-10.

[2] Gaieski DF, Edwards JM, Kallan MJ, Carr BG. Benchmarking the incidence and mortality of severe sepsis in the United States. Crit Care Med. 2013; 41:1167–74.

[3] Kumar A, Haery C, Paladugu B, et al. The duration of hypotension before the initiation of antibiotic treatment is a critical determinant of survival in a murine model of Escherichia coli septic shock: association with serum lactate and inflammatory cytokine levels. J Infect Dis. 2006; 193:251–8.

[4] Seymour CW, Gesten F, Prescott HC, et al. Time to Treatment and Mortality during Mandated Emergency Care for Sepsis. N Engl J Med. 2017 June 08; 376(23): 2235–44.

[5] Paoli CJ, Reynolds MA; Sinha M. Epidemiology and Costs of Sepsis in the United States—An Analysis Based on Timing of Diagnosis and Severity Level. Crit Care Med. 2018 Dec 46(12):1889-97.

[6] Rhodes A, Evans LE, Alhazzani W, et al. Surviving Sepsis Campaign: international guidelines for management of sepsis and septic shock: 2016. Intensive Care Med. 2017; 43:304–77.

[7] Rhee C, Gohil S, Klompas M. Regulatory mandates for sepsis care — reasons for caution. N Engl J Med. 2014; 370:1673–6.

[8] Ferrer R, Martin-Loeches I, Phillips G, et al. Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program. Crit Care Med. 2014; 42:1749–55.

[9] 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.

Address for correspondence.

Reza Firoozabadi
Philips Healthcare
3000 Minuteman Rd, Andover, MA 01810
reza.firoozabadi@philips.com