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. Our team (AlgTeam, ranking 39) submitted nine entries for evaluation on the subset of the hidden data and selected the entry with highest utility score which resulted in the final utility score of 0.24 on the full test dataset.

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. A utility score was defined by the challenge organizers, rewarding early predictions of sepsis and penalizing too early/late/failing prediction of sepsis or false sepsis prediction in a non-sepsis patient. For detailed description of the challenge, refer to the publication by the challenge organizers [9].

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 bootstrap-aggregated decision trees with a highly unbalanced misclassification cost function was developed to predict sepsis for each time sample of patient features. 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 training 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. Multi-feature record of each patient consists of several samples of features typically collected every hour. Features were preprocessed in several steps including the outlier elimination, combination of the correlated variables, missing value imputation, and standardization.

The preprocessed records were then split into training

 ISSN: 2325-887X DOI: 10.22489/CinC.2019.023

and validation datasets. A tree-bagger classifier was trained using the training dataset and the important features and the hyperparameters were selected in an iterative approach until the best score was achieved. The classifier was validated by validation dataset and the model was submitted to evaluate the score of a subset of the hidden dataset. The classifier with the best utility score was selected for evaluation of the final utility score on the full test dataset.

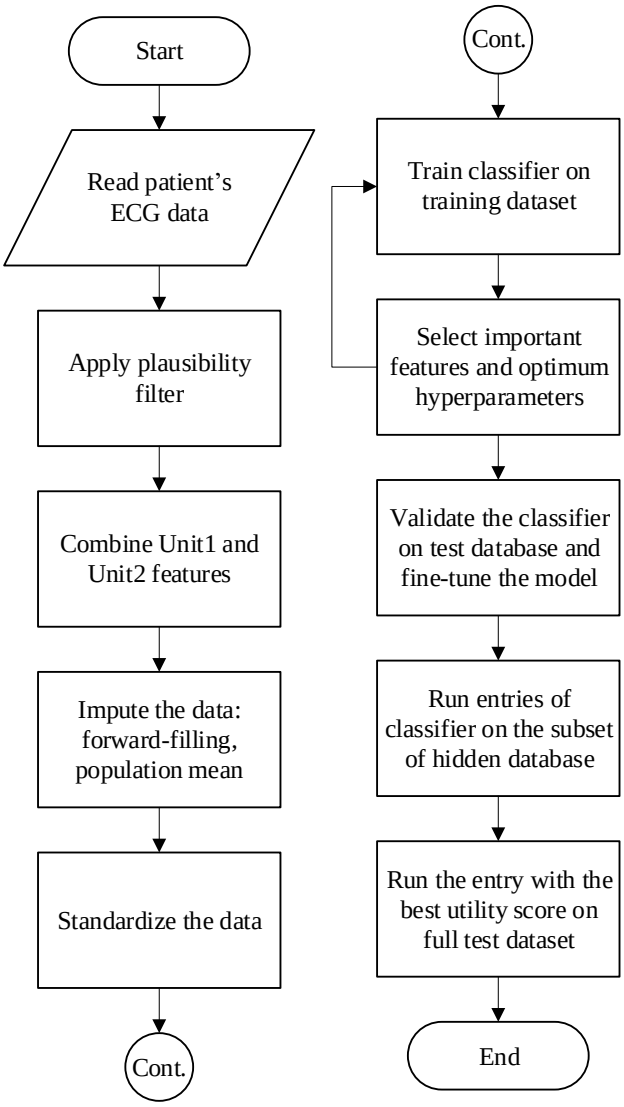

Figure 1. Block diagram of the algorithm consisting of thje data preprocessing (left branch) and the evaluation procedure (right branch).

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 to the challenge participants and used in the algorithm training. Datasets A and B contain 20,336 and 20,000 patient records, respectively. The third dataset was hidden and used for scoring and evaluation of the algorithm for the final utility score. The complications were the highly unbalanced data (only 1.8% of the patient records showed sepsis) and high number of missing values (up to 99.8% in some features) in public databases.

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 the features in each group, their missing percentage, preprocessing information, and the statistics.

2.2.1. Data Preprocessing

The patient records were preprocessed before being used in development of the classifier model. The first step was applying a plausibility filter to each feature. A range of valid values for each feature was defined based on its actual distribution and the knowledge in the literature. Any value outside this range was assumed as 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 since the patient was either in MICU or SICU. In case of missing values, MICU (Unit1) was assumed.

There is large number of missing values in some features. This percentage ranges 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 a value was available in past. The remaining missing values with no previous 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. The statistical values for each feature after preprocessing are shown in Table 1.

Figure 2 shows an example of data from a patient after preprocessing the features and combining Unit1 and Unit2 features. The preprocessed features are the imputed and standardized time series samples. In this example there are 39 features varying in a 54-hour time interval. Most features are missing in the early hours of data collection and are imputed with the population mean values. As observed, some features are collected rarely and imputed by previous values while some are collected frequently.

Table 1. Features in each group and their mean, median, percentage of missing values, and the low and high values in plausibility range after preprocessing.

	Features	Missing (%)	Plausibility		Mean	Median	SD
			Low	High			
Vital signs	HR	9.9	10	300	84.6	83.5	17.3
	O2Sat	13.1	60	100	97.2	98	2.7
	Temp	66.2	32	42.2	37	37	0.8
	SBP	14.6	40	280	123.8	121	23.2
	MAP	12.5	0	300	82.4	80	16.3
	DBP	31.3	20	130	63.7	62	13.6
	Resp	15.4	5	60	18.8	18	5
	EtCO2	96.3	0	150	33	33	8
Laboratory values	BaseExcess	94.6	-20	20	-0.7	0	4.2
	HCO3	95.8	0	50	24.1	24	4.4
	FiO2	91.7	0	1	0.5	0.5	0.2
	pH	93.1	6	8	7.4	7.4	0.1
	PaCO2	94.4	0	200	41	40	9.3
	SaO2	96.5	0	100	92.7	97	10.9
	AST	98.4	0	400	64.5	35	73
	BUN	93.1	0	500	23.9	17	20
	Alkalinephos	98.4	0	250	82.8	71	43
	Calcium	94.1	0	20	7.6	8.3	2.4
	Chloride	95.5	75	145	105.8	106	5.8
	Creatinine	93.9	0	10	1.4	0.9	1.4
	Bilirubin_direct	99.8	0	50	1.8	0.4	3.7
	Glucose	82.9	0	1000	136.9	127	51.3
	Lactate	97.3	0	100	2.6	1.8	2.5
	Magnesium	93.7	0	10	2.1	2	0.4
	Phosphate	96.0	0	12	3.5	3.3	1.4
	Potassium	90.7	1	10	4.1	4.1	0.6
	Bilirubin_total	98.5	0	50	2.1	0.9	4.3
	TroponinI	99.0	0	200	8	0.3	22.7
	Hct	91.1	10	70	30.8	30.3	5.5
	Hgb	92.6	2	22	10.4	10.3	2
	PTT	97.1	0	250	41.2	32.4	26.2
	WBC	93.6	0	50	11.2	10.3	5.4
	Fibrinogen	99.3	0	800	280.2	248	137.5
	Platelets	94.1	5	1500	196	181	103
Demographics	Age	0.0	0	150	62	64	16.4
	Gender	0.0	0	1	0.6	1	0.5
	Unit1	39.4	0	1	0.5	0	0.5
	Unit2	39.4	0	1	0.5	1	0.5
	HospAdmTime	0.0	none	none	-56.1	-6	162.3
	ICULOS	0.0	1	none	27	21	29

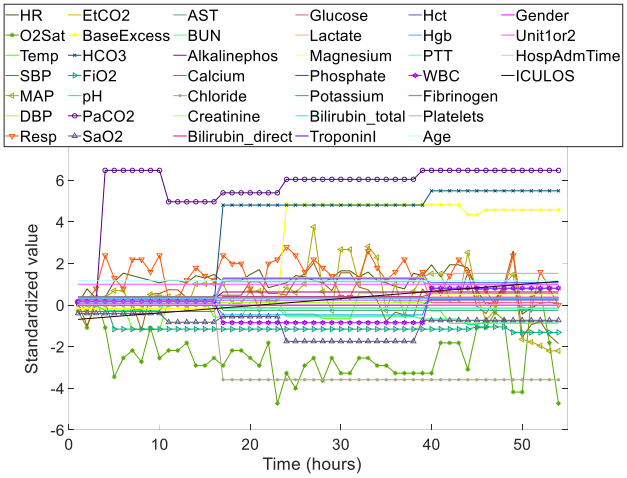

Figure 2. An example of a patient's preprocessed data with 39 features varying in the 54-hour time interval.

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 trained model accepts the important features as input. The ensemble consists of 100 decision trees. Due to the highly unbalanced nature of the data, a misclassification cost ratio of 1 to 37 was defined for no-sepsis versus sepsis. Maximum number of splits is set to 100 with minimum leaf size of 3.

Figure 3 displays an example of an ensemble of 100 decision trees.

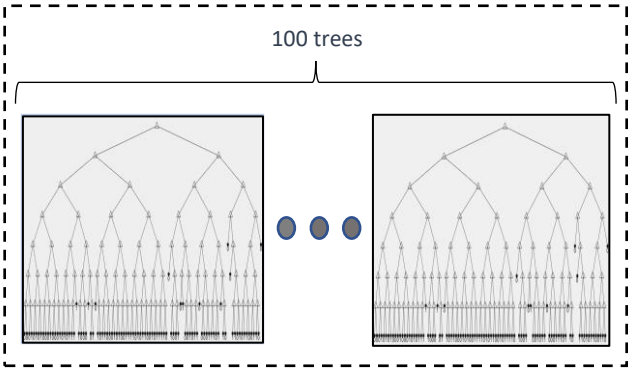

Figure 3. Example of one decision tree.

2.3.1. Training and Feature Selection

Our predictive model was trained using the features selected from dataset A. Important features and optimized hyperparameters were selected in an iterative optimization approach of the utility score. A total of 15 features were selected: HR, Temp, MAP, Resp, BaseExcess, FiO2, BUN, Calcium, Creatinine, Hct, WBC, Platelets, Unit1or2, HospAdmTime, ICULOS.

Multi-feature sample of the data in each patient at each time was used as the single input to the classifier, resulting in a single binary output of sepsis or no-sepsis. Up to three initial samples in each patient with excessive missing values were discarded from analysis. Dataset B was used to validate and fine-tune the optimized model.

3. Results

Our team participated in the challenge with the team name *AlgTeam*. We submitted one successful entry in the unofficial phase and a total of nine successful entries in the official phase. We selected the entry with the highest utility score as the final entry for running on the full test dataset. The final utility score on the full hidden dataset was 0.24 ranking 39. Execution time was 14 hours and 23 minutes on test set A.

Table 2 shows the detailed results of running the model in our selected entry on each test dataset (A, B, or C), including the utility score, AUROC, AUPRC, accuracy, and the F-measure.

4. Discussion and Conclusions

In order to compare the performance of our algorithm with the other teams participated in the challenge, we calculated the average results for all ranked teams where their average utility score was 0.185 on the full test set (Table 3).

Compared to the other teams' average results, our algorithm showed high utility score on both datasets A and B, while its score on dataset C was close to the average. However, our utility score for dataset B was much lower than dataset A. One reason is that high correlation was observed between the results from the official submissions and the dataset A, hence we focused on training our algorithm by dataset A only and validated it by dataset B. If we had the prior knowledge that the official phase test subset was a subset of the dataset A and not from another hospital, and the full test dataset includes a subset of dataset B as well as another hospital, we would have trained and validated our algorithm on both datasets A and B. We also trained our algorithm using the single time samples of the features and did not regard the correlation with previous samples. This have probably caused the low score from dataset C. Adding the handcrafted correlation features or using a sequence model such as LSTM may improve the performance.

Also compared to the other teams' average results, our AUROC measures for all datasets were higher, our AUPRC were higher for dataset A and similar for the other two datasets, our accuracy was higher for all datasets, and our F-measure was higher for datasets A and B, but comparable for dataset C.

Table 2. Results of running our selected model.

	Score	AUROC	AUPRC	Accuracy	F-measure
Set A	0.335	0.764	0.084	0.871	0.139
Set B	0.268	0.768	0.055	0.912	0.118
Set C	-0.226	0.741	0.033	0.754	0.039

Table 3. Other teams' average results.

	Score	AUROC	AUPRC	Accuracy	F-measure
Set A	0.267	0.586	0.058	0.819	0.108
Set B	0.211	0.598	0.053	0.859	0.098
Set C	-0.219	0.581	0.035	0.762	0.041

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