Sepsis Prediction Model Based on Vital Signs Related Features

Vytautas Abromavičius[1], Artūras Serackis[1]

[1] Vilnius Gediminas Technical University, Vilnius, Lithuania

Abstract

In this paper, we present our solution for early detection of sepsis by joining the PhysioNet/Computing in Cardiology Challenge 2019. Our proposed algorithm uses three different models for sepsis prediction. The model is selected according to the time the patients have already spent in the intensive care unit. The first model uses 64 features and s applied if the patient stays in ICU for the first 9 hours. Second and third models use more advanced 111 features. The second prediction model is activated if the patient stays for more than 9 hours. The third prediction model is activated for more extended stays if the patient stays for more than 60 hours. Inspection of the data, and consultation with personnel of the local ICU led us to believe that the time patients spent in the ICU or were hospitalized is an essential indicator for the risk of developing sepsis. During the longer stays in hospital number of intravenous measurements and other procedures increases, increasing the risk of blood infection. Therefore, feature extraction in our algorithm was based on these metrics. The best-received score on A and B datasets with the models, trained using Gentle Boosting on a training set with ADASYN balancing was 0.5955, and the best score with the models, trained on a dataset with randomly removed samples was 0.7616. However, the official score on our best algorithm was only 0.036.

1. Introduction

Sepsis is a syndrome of physiologic, pathologic, and biochemical abnormalities induced by infection [1]. The conservative estimates indicate that sepsis is a leading cause of mortality and critical illness worldwide [2, 3]. World Health Organization concerned that sepsis continues to cause approximately six million deaths worldwide every year, most of which are preventable [4, 5]. In their study the Department of Health in Ireland reported that survival from sepsis-induced hypotension is over 75% if it is recognized promptly, but that every delay by an hour causes that figure to fall by over 7%, implying that the mortality increases by about 30% [6].

In this paper, we present our solution for early detection of sepsis by joining the PhysioNet/Computing in Cardiology Challenge 2019. The participants were challenged to predict sepsis six hours before the clinical prediction. The sepsis in this challenge was defined according to the Sepsis-3 guidelines: a two-point change in the patient's Sequential Organ Failure Assessment (SOFA) score and clinical suspicion of infection (as defined by the ordering of blood cultures or IV antibiotics) [1]. Our open-source algorithm works on clinical data provided on a real-time basis by giving a positive or negative prediction of sepsis for every single hour.

Data used in the competition was sourced from ICU patients in three separate hospital systems. Data from two hospital systems were publicly available and was used to create and test our algorithm. Data includes eight vital signs up to 26 laboratory values, information about the age, gender, hours between hospital admit and ICU admit, hours since ICU admit.

Our proposed algorithm was scored on a censored data set, dedicated for scoring and using utility function that rewards early predictions and penalizes late predictions as well as false alarms.

In the absence of clearly defined and highly accurate diagnostic tools, the ICU physicians rely on their own clinical skill set and experience to diagnose sepsis. The experience among clinical features also includes prediction of the possible source of infection [7]. It is common for some sepsis patients to have variations of temperature during four or six-hour period, which may become higher than 38 or lower than 36. High temperatures may be followed by the sudden arterial blood pressure drop. The risk of sepsis also is related to the time spent in ICU. Also, each intervention increases this risk.

According to the labels in the public datasets, the sepsis for the ICU patients may be developed during several hours after admission to ICU or much later. Taking into account that the risk of sepsis development increases with time spent in ICU, we decided to treat vital signs, laboratory values, and other clinical data differently for different length of stay in ICU.

In the following sections we describe data preparation techniques we have applied before feature extraction. Then we explain which indicators we used to extract features.

Next we present classifiers which we have tested together with disscussion on the received investigation results.

2. Materials and Methods

The sepsis prediction algorithm we present in this paper is based on the selection of the different set of features according to the time spent in ICU. The duration of the first period is selected for up to 9 hours. A specific set of features and classifier is applied during this first period. Another classifier uses the features dedicated for the second period if the duration of stay in ICU is 10 hours or more, but less than 61 (approximately two and a half-day). If the patient stays in the ICU for 61 hours or more, we introduce the third classifier and dedicated features.

2.1. Data Preparation

To develop our algorithm and to train models we have used the initial dataset of around 5000 records which was available after the PhysioNet/Computing in Cardiology Challenge 2019 was initiated and both training sets: training set A (20,336 subjects) and B (20,000 subjects). Each dataset contains several measurements for each patient and a sepsis label.

The training datasets contains highly unbalanced data. Only 7.08% (3211 from 45336) records were from patients with sepsis confirmed. Such situation required specific approach in order to train a classifier on this data. In order to balance the data, we have used Adaptive synthetic sampling (ADASYN) algorithm [8].

Taking into account the fact that the models in our algorithm should work on a real-time basis, we have duplicated dataset entries to create an input-output mapping on an hourly basis. Measurements made at the first hour were used with the labeled situation at that hour. Next, these measurements were added to those which were made during the next hour and used as a separate entry with the situation label (desired output), market after two hours. Such duplication for a patient, who spent in ICU 30 hours gives 30 entries in total for the dataset, prepared to train models of our algorithm.

Our algorithm, proposed in this paper, used three differently trained models. Prepared for model training entries were divided into three subsets. The first subset was with entries from 1 hour to 9 hours spent in ICU. For the patients who spent in ICU for more than 9 hours, we took records from the first 9 hours, spent in ICU. The second subset was with entries from 10 to 60 hours spent in ICU. The third subset included entries with duration over 61 hours spent in ICU. To make things easier, we called these subsets *short*, *medium*, and *long*.

2.2. Feature Extraction for Short Stay Period

Making sepsis prediction on a *short* subset differs a little bit compared to the subsets which represent longer stays in the ICU. Several measurements do not give a possibility to rely on features with advanced estimation. Therefore, we have used simple tools, such as mean, median, entropy, standard error, age, and hours between hospital admit and ICU admit.

The feature vector of 64 elements was used as an input to the model, trained on a *short* dataset and used for prediction if the patient is in the ICU for a period not longer than 9 hours. First 34 elements in the feature vector are the mean values, calculated from the first 34 types of measurements in the data file. It includes all eight vital signs and laboratory values if present. It is essential to notice that the mean value is calculated, not taking into account NaN values (missing records) in the patient's data file. If during the analyzed period, there were no measurements of a specific type, we put 0 in the feature vector.

The second 7 elements in the feature vector are the median of the measurements in the data file, calculated for the following vital signs: HR Heart rate (beats per minute), O2Sat Pulse oximetry (%), Temp Temperature (Deg C), SBP Systolic BP (mm Hg), MAP Mean arterial pressure (mm Hg), DBP Diastolic BP (mm Hg), Resp Respiration rate (breaths per minute). For these vital signs, we have also calculated other features: the Shannon entropy, stored from 69 to 75 in the feature vector; the kurtosis, stored as the following seven values in the feature vector; and standard error. The last two elements in the feature vector are the age of the patient and the hours between hospital admit and ICU admit.

2.3. Feature Extraction for Medium and Long Stay Periods

For the *medium* and the *long* subsets, we used a vector with 111 elements as an input to the models, trained for each subset individually. However, the feature selection was a bit more complicated, than in the case with *short* data subset.

First difference with the features, used for *short* data subset is that not all measurements were used to calculate features for the model. Using statistical analysis of variance (ANOVA), we selected only the following record types: HR, Temp, SBP, DBP, Resp, EtCO2, Calcium, Lactate, and Hct. For these nine types of measurements, we have calculated: mean value, Shannon entropy, kurtosis, standard error. These features were used as the first 36 elements in the feature vector. We also added the age of the patient and the hours between hospital admit and ICU

admit as 37^{th} and 38^{th} feature respectively.

Second, to calculate the 39^{th} feature, we selected a set of measurements, which usually (more frequently) were taken separately: FiO2, pH, PaCO2, SaO2, Calcium, Glucose, Lactate. This set was used to calculate a specific feature for a model. To calculate this feature, we took the number of measurements for each measurement type and divided it by the total length of stay in ICU. As a feature, we took an average of these seven resulting values.

Selected nine types of measurements also were used to calculate an additional set of features, which, according to our idea, should describe the current situation of the patient. We have selected 7 hour duration time frame and calculated the following features: standard deviation, mean, difference between last value and the maximum value, which accrued during selected time frame, difference between maximum and minimum values, difference between last two measurements, standard deviation of differences between measurements, maximum difference between measurements over last 7 hours.

2.4. Classification

In order to speed-up the training of the models, we applied input data standardization. The mean and standard deviation coefficients were estimated using the measurements taken from the whole training dataset. Those estimated coefficients are now included in our algorithm as constant values.

For our algorithm, presented in this paper, we have used three trained models. All models were based on decision tree classifier, trained using Gentle Adaptive Boosting ensemble learning algorithm [9]. Also, we have trained the models of our algorithm using Random undersampling boosting [10], however, the performance was very low.

3. Results

There were two types of models trained and tested on a training dataset: models based on Gentle Boosting [9] and models based on RUS boosting [10]. The final algorithm uses the same type of the model for all three subsets: *short*, *medium*, and *long*.

Table 1. Sepsis prediction results using differently trained classification models.

Model	AUROC	AUPRC	Accuracy	F-measure	Utility
GBMod1	0.2033	0.0059	0.9583	0.2782	0.4351
GBMod2	0.0322	0.0009	0.9554	0.3677	0.7616
GBMod3	0.1129	0.0034	0.9142	0.1964	0.5184
GBMod4	0.0979	0.0029	0.9370	0.2592	0.5955

Since we have trained our models on differently divided and pre-processed datasets, we decided to show some best-received results in Table 1.

Three models of the algorithm, used in the algorithm version named GBMod1, were trained on the unbalanced dataset, where the number of examples with sepsis situations was only 1% comparing to 99% of situations when sepsis should not be predicted (indicated). The models were based on Gentle Boosting Ensemble method.

The models used in algorithm GBMod2 also used Gentle Boosting Ensemble method, but a significant amount of non-sepsis situations were randomly removed from the training data. The resulting training dataset had 20% of sepsis labeled situations and 80% of non-sepsis situations. Such an approach for training showed the best results and achieved a score of 0.7616.

GBMod3 and GBMod4 types models were trained on a balanced training dataset. The balancing was made using ADASYN algorithm. Difference between these types of modes is that GBMod3 model for *medium* subset was trained using a Maximum number of splits equal to 30 and GBMod4 model for the same subset was trained using a Maximum number of splits equal to 75. Models with a higher number of splits performed better.

RBMod1 is an alternative type of classification model. Models of this type are based on RUS Boosting algorithm and were trained on balanced training data. We also set the maximum number of splits to 75.

4. Conclusions

Our approach, presented in this paper, was created using the traditional way by trying to find and select a set of features, which can give the best classification performance.

We have selected Decision tree classifiers working on 64 or 111 features, depending on how long the patient already stays in ICU. Training of the models using ensemble methods showed that enabled PCA reduces the number of feature from 111 to 65. However, the accuracy decreases by almost 8%.

Balancing the dataset using ADASYN did not show better results compared to the random selection of fewer members from the class with a higher number of elements for training. The best-received score on A and B datasets with the models, trained using Gentle Boosting on a training set with ADASYN balancing was 0.5955, and the best score with the models, trained on a dataset with randomly removed samples was 0.7616. However, the official score on our best algorithm was only 0.036.

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). Jama 2016;315(8):801–810.

[2] Vincent JL, Marshall JC, Ñamendys-Silva SA, François B, Martin-Loeches I, Lipman J, Reinhart K, Antonelli M, Pickkers P, Njimi H, et al. Assessment of the worldwide burden of critical illness: the intensive care over nations (icon) audit. The lancet Respiratory medicine 2014; 2(5):380–386.

[3] Fleischmann C, Scherag A, Adhikari NK, Hartog CS, Tsaganos T, Schlattmann P, Angus DC, Reinhart K. Assessment of global incidence and mortality of hospital-treated sepsis. current estimates and limitations. American journal of respiratory and critical care medicine 2016;193(3):259–272.

[4] Reinhart K, Daniels R, Kissoon N, Machado FR, Schachter RD, Finfer S. Recognizing sepsis as a global health priority—a who resolution. New England Journal of Medicine 2017;377(5):414–417.

[5] (WHO) WHO, et al. Improving the prevention, diagnosis and clinical management of sepsis. Geneva WHO 2017;.

[6] of Health D. Sepsis management national clinical guideline.

[7] Klouwenberg PMK, Ong DS, Bos LD, de Beer FM, van Hooijdonk RT, Huson MA, Straat M, van Vught LA, Wieske L, Horn J, et al. Interobserver agreement of centers for disease control and prevention criteria for classifying infections in critically ill patients. Critical care medicine 2013;41(10):2373–2378.

[8] He H, Bai Y, Garcia EA, Li S. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). IEEE, 2008; 1322–1328.

[9] Friedman J, Hastie T, Tibshirani R, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics 2000;28(2):337–407.

[10] Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A. Rusboost: Improving classification performance when training data is skewed. In 2008 19th International Conference on Pattern Recognition. IEEE, 2008; 1–4.

Address for correspondence:

Artūras Serackis
Naugarduko g. 41-413, Vilnius, LT-03227, LITHUANIA
arturas.serackis@vgtu.lt