Early Prediction of Sepsis from Clinical Data Using single Light-GBM model

Soufiane Chami[1], Naima Kaabouch[1] , Kouhyar Tavakolian[1]

[1] Dept. of Electrical Engineering & Computer Science
University of North Dakota, Grand Forks, ND, United States

Abstract

— Sepsis is a severe medical condition caused by body's extreme response to an infection causing tissue damage, organ failure and even death. The emergence of advanced technologies like Artificial Intelligence and machine learning, allowed faster exploration of advanced way to recognize sepsis cases. In this paper, we will present two main approaches that has been tested using medical data. First is the combination of survival analysis and neural networks, and the second is based on booting trees. Our best single model got 0.172 for normalized utility score and 0.272 for AUROC.

1. Introduction

Sepsis is a severe medical condition caused by body's extreme response to an infection causing tissue damage, organ failure and even death. Every 3-4 seconds, at least one person dies because of sepsis worldwide. Sepsis affect about 1.6 million American yearly [1]. As a leading cause of death in the US hospitals, sepsis costs the US about 24 billion USD, 6.2% of the total hospital costs in 2013. [2].

The early detection of sepsis has proven to be a key factor in increasing the efficiency of antibiotic treatment for septic patients. Related studies [3] demonstrated that early detection will prevent 80% of death cases caused by sepsis. On the other hand, it was reported that the sepsis mortality significantly increases with the length of stay of the septic patient in the hospital. In order words, delayed recognition of sepsis exacerbates the risk of death of a septic patient by 7.6% every hour [4].

Furthermore, with the emergence of more advanced technologies, there is a significant amount of Electronic Health Records (EHRs) that became available. The EHRs are a systematic collection of data used as health indicators of a patient. The growing availability of the EHRs brought so much interests and opportunities to think of more advanced predictive models to early recognize septic patients. The electronic health records can be dynamic changing over time (time series) like heart rate or blood pressure. Also, they can be static like demographic information such as Age and gender.

The clinical definition [5] of systemic inflammatory response (SIRS) to infection, specifies four conditions which only two of them are sufficient to trigger an alert of sepsis:
1. Temperature of $> 38^o$C or $< 36^o$C
2. White blood cell count of $> 12,000$ per ml or $< 4,000$ per ml, or $> 10\%$ immature (band) forms
3. Heart rate of ¿90 beats per minute
4. Respiratory rate of > 20 breaths per minute or partial pressure of CO2 of $<$32 mmHg

In the following sections, the problem of the PhysioNet challenge of this year and the data provided are discussed. Then, the proposed model to solve this challenge is presented.

2. Physionet Challenge and Data

Saving septic patients' lives comes with a combination of efforts from different perspective of sepsis. While more advanced instruments allow closer and more accurate follow-up of the septic patient situation. They only can tell about the current situation of a patient in a potential risk of sepsis. [6]

2.1. Problem definition

In most of the time when a patient arrive to the confirmed sepsis shock stage, it most probably too late for us to save their lives. Launching the antibiotic culture take time and is less effective when a patient is in serious stage of sepsis that become pretty evident.

The instrument tells what have already happened to the patient but alone they cannot tell a lot about the future development of the patient situation. This later information is more relevant for saving infectious patient from death of sepsis.

An effective treatment of sepsis is performed through different aspect and numerous sides. From one side, there

is the instrumentation part, and on the other hand, there is the computational part.

It's worth noting that early detection and recognition of sepsis and antibiotic treatment are critical key factor of improving sepsis outcomes [7].

Each time we detect sepsis patient one hour earlier, we get more 4-8% chance to save the patient's life. With the classical tools of data analysis in the last decade, scientist still face a lot of limitations to detect sepsis early enough.

With the emergence of machine learning techniques, a lot research has been made to improve the current detection tools. [1, 8, 9].

2.2. Related work

The early detection of sepsis has proven to be a key factor in increasing the efficiency of antibiotic treatment for septic patients. Related studies [10] demonstrated that early detection will prevent 80% of death cases caused by sepsis. On the other hand, it was reported that the sepsis mortality significantly increases with the length of stay of the septic patient in the hospital.

The complexity of sepsis is still high and depends on so many factors. There were so many studies and attempts to propose a predictive model for this purpose, but they are facing some challenges.

Chen Lin, in his work [7, 11], used dynamic EHRs and proposed a generic framework of deep learning to detect septic patients 5 hours earlier. The framework consisted of two experimental setups: one with Convolutional Neural Network (CNN) added before Long Short-Term Memory (LSTM) to extract patterns from time series features (e.g. EHRs). This component will be connected later with a fully connected network. This later component makes use of the static features and extract local characteristics and dependencies in this type of EHRs.

LSTM is very effective in extracting patterns from long sequences. It's always considered as strong candidate to handle date with temporal structures like time series signals, videos, or text data... etc.

In fact, LSTM [1] is neural network with a memory component that save the previous inputs and use them along with the current inputs. This is a very efficient strategy to capture the dynamic dependencies of the EHRs for sepsis patients. It will help to detect the physiological deterioration in the data earlier than a classical structure of neural network with no memory.

In addition, the CNN [8] is more known for image data rather than temporal data. This structure has achieved great results on many hard topics such as object detection. There was a work made by Krizhevsky where the temporal data was converted into image format and fed to a CNN. This achieve interesting results and shown promising potential of applied CNN on time series.

3. Method

3.1. Data Features

There are more than 40 health variables used to track the health situation of the patients in this challenge. The 40 columns are classified into three classes: vital signs, lab test, and static variables.

3.1.1. Vitals signs

Most of these features concern the main screening signal that would reflect continuous situation of the patient health. There are 8 vital signs provided in the data and we can cluster them into three main categories: ECG signals, Pulse signals, and Temperature.

- **ECG signals:** The main columns related are the heart rate, systolic blood pressure and diastolic blood pressure. The severity of sepsis is very correlated to the number of beats per minute. Many studies have shown that the more sepsis get worse, the more we observe ECG abnormalities. This can be explained by loss of excitability in cardiac tissue during the sepsis.
- **Pulse signals:** Respiration rate, O2Sat, and EtCO2 (End tidal carbon dioxide)
- **Temperature:** Symptoms of sepsis include: a fever above 101°F (38°C) or a temperature below 96.8°F (36°C) heart rate higher than 90 beats per minute. So, tracking the temperature is very important for the early detection task. In fact, the predictive model uses a certain number of the previous data points from the current time and try to detect any significant shift on temperature (drop or rise).

3.1.2. Lab tests

The role of laboratory test is very significant for early detection of sepsis. Since the main definition of sepsis is the body's systemic inflammatory response to a bacterial infection [12]. The spread of bacteria in the blood (bacteremia) make it a big indicator of sepsis infection. About 25 lab tests have been provided in the data for sepsis detection. These features include: White blood cells count, Blood urea nitrogen, Lactic acid, partial thromboplastin time, Leukocyte count, Platelets.

3.1.3. Demographics features

There were some studies that have shown that sepsis mortality has a little thing to do with demographics. Especially age and length of stay. In our current model, we still find this is not the case. More results will be shown on the next paragraphs. [13]. The main demographic features in the data are: Gender, Age, and length of stay.

3.2. Data Pre-processing & Features

Before the learning step, the clinical dataset we have for the challenge contain a lot of inconsistencies. For our model we have performed many preprocessing techniques to ensure the consistency of the data and create new features. The processing pipeline can be described as follow:

(1) Handling missing values: Several lab tests features have up to 90% of missing values. The data resolution is hourly based, and it's difficult to perform lab experiments every hour for every patient. Which explain the high ratio of missing values in lab test columns. However, the values in these columns are very important and removing them would be the best idea:

1. We have performed interpolation which fill the missing values with the mean of the two consecutive non-missing values.
2. The remaining missing values are filled backward and then forward.

(2) Lag features with different sliding windows: mean of the last [1,.., 9] hours. The purpose of this features is to capture the long- and short-time dependencies.

(3) Data binning: In order to reduce some variance in the signal columns. We create new features besides to the original signals that would aggregate the signal value in ranges. We have based the data binning on the max and min. There is an option to define range limits using Random Forest, but we didn't explore it in this work.

(4) Count Encoding of Categorical features:: the main target of this are the demographics features such as: age, gender. . . etc.

3.3. Model Overview

While some research works [11,12] showed a great ability to predict sepsis in infectious patients. It was still limited since many of these simulations could not predict earlier than 6 hours. As this time frame increases, the accuracy decreases a lot.

In this section, we are going to outline two approaches that we continue improving in order to use the features prementioned earlier, to predict sepsis patient no earlier than 12 hours and no later than 6 hours.

3.3.1. Weibull Time-To-Event- RNN model

Early detection is the typical application of Survival Analysis (SA); it has been used in medicine and early detection of catastrophic events.

However, SA despite being able to fit perfectly and handle the censored nature of data used to estimate Time-To-Event (TTE) problems, it still remains a statistical modelling approach and not able to keep up with the advanced approaches of deep learning.

Likewise, the fundamental idea of deep learning architectures still does not take into consideration the censorship aspect of the data as is needed TTE problems. As a result, deep learning itself would not be able to provide high performance on such problems. In this paper, this model couldn't show a huge performance in the challenge. But we believe it's very promising and deserve to be mentioned. It will be the subject improvement in future work.

In Weibull Time-To-Event- RNN model (WTTE-RNN) [14] model we assume that TTE variable follow Weibull distribution governed by two parameters: alpha and beta. The idea is to estimate the TTE distribution instead of directly estimating the TTE variable. This approach was not very successful, it's still vague how the algorithm is learning in the data. We needed to show more clarity in the model and that's why we thought about boosting trees. Light-GBM was the best candidate.

3.3.2. Light-GBM

This approach is a gradient boosting learning framework that uses tree-based learning algorithm. The difference between light-GBM and other tree-based algorithms is that it grows tree leaf-wise while other algorithms grows level-wise.

It will choose the leaf with max delta loss to grow. When growing the same leaf, Leaf-wise algorithm can reduce more loss than a level-wise algorithm.

3.3.3. Cross Validation

In order to solve the class unbalance in the training, we went for a cross validation of 5 stratified folds. The index sampling made patient wise not data point wise. This later would cause a huge data leak and biased training.

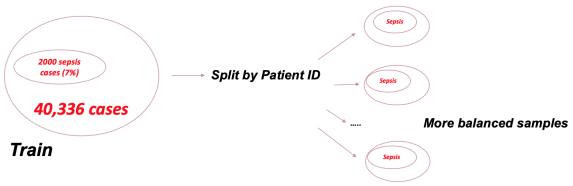

Figure 1. Cross validation

3.4. Results

In this section, we are going to present the results of our best submission. The improvement of WTTE-RNN will discussed in future work. Light-GBM: The results of this approach are more interesting. We not only got a great classifier, but we also learned those variables that contribute the most in the prediction of sepsis. With AUC

validation score (83%), we have learned some importance features that contribute to the prediction. On the official results, we got 0.272 for AUROC, which is the inverted classification score. The primary results report that some demographics like Age and Hours between hospital admit and ICU admit are very significant in predicting sepsis patient. Furthermore, the lab test that concern the cell count are very significant. For instance, white blood cells and Platelets are very helpful features. The Figure below gives more details about features importance.

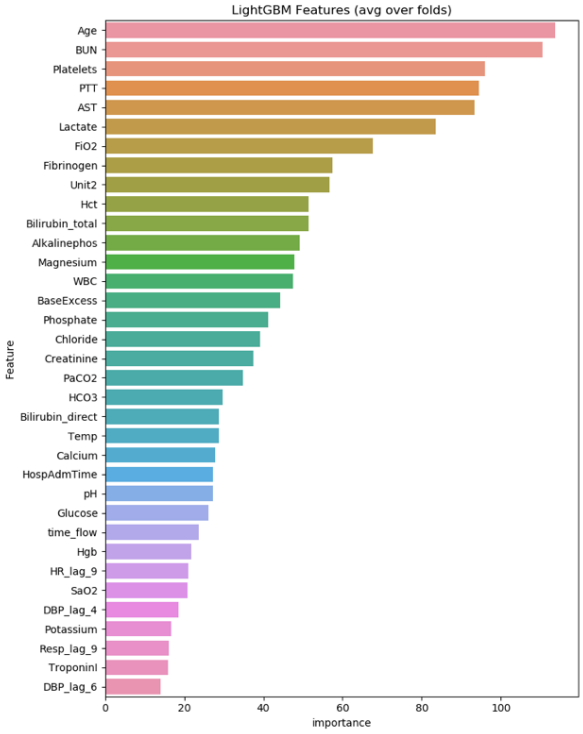

Figure 2. Feature Importance provide better model clarity.

References

[1] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation 12 1997;9:1735–80.

[2] Umscheid C, Betesh J, VanZandbergen C, Hanish A, Tait G, Mikkelsen M, French B, Fuchs B. Development, implementation, and impact of an automated early warning and response system for sepsis. Journal of Hospital Medicine 09 2014;10.

[3] Kumar A, Roberts D, Wood K, Light B, Parrillo J, Sharma S, Suppes R, Feinstein D, Zanotti S, Taiberg L, Gurka D, Kumar A, Cheang M. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Critical care medicine 07 2006;34:1589–96.

[4] Frost R, Newsham H, Parmar S, Gonzalez-Ruiz A. Impact of delayed antimicrobial therapy in septic itu patients. Critical Care 09 2010;14:1–2.

[5] Hotchkiss R, Moldawer L, Opal S, Reinhart K, Turnbull I, Vincent JL. Sepsis and septic shock. Nature Reviews Disease Primers 06 2016;2:16045.

[6] MA R, Josef C JR, Shashikumar SP WM, Nemati S CG, A S. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. Machine Learning and Applications An International Journal 09 2019;in press.

[7] Zhang Y, Lin C, Chi M, Ivy J, Capan M, Huddleston J. Lstm for septic shock: Adding unreliable labels to reliable predictions. 12 2017; 1233–1242.

[8] Krizhevsky A, Sutskever I, Hinton G. Imagenet classification with deep convolutional neural networks. Neural Information Processing Systems 01 2012;25.

[9] Futoma J, Hariharan S, Heller K. Learning to detect sepsis with a multitask gaussian process rnn classifier 06 2017;.

[10] Kumar A, Roberts D, Wood K, Light B, Parrillo J, Sharma S, Suppes R, Feinstein D, Zanotti S, Taiberg L, Gurka D, Kumar A, Cheang M. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Critical care medicine 07 2006;34:1589–96.

[11] Lin C, Zhangy Y, Ivy J, Capan M, Arnold R, Huddleston J, Chi M. Early diagnosis and prediction of sepsis shock by combining static and dynamic information using convolutional-lstm. 06 2018; 219–228.

[12] Fan SL, Miller N, Lee J, Remick D. Diagnosing sepsis – the role of laboratory medicine. Clinica Chimica Acta 07 2016;460.

[13] Menezes B, Araújo F, Amorim F, Santana A, Soares F, Souza J, Araújo M, Santos L, Rocha P, Gomes M, Neto O, Júnior P, Amorim A, Biondi R, Ribeiro R. Comparison of demographics and outcomes of patients with severe sepsis admitted to the icu with or without septic shock. Critical Care 11 2013;17:P48.

[14] Cawley R, Burns D. Analysis of wtte-rnn variants that improve performance. Machine Learning and Applications An International Journal 03 2019;35–47.

Address for correspondence:

Soufiane CHAMI
Biomedical Engineering Research Complex - UND
soufiane.chami@und.edu