An Ensemble Machine Learning Model For the Early Detection of Sepsis From Clinical Data

Mengsha Fu[1], Jiabin Yuan[1], Menglin Lu[2], Pengfei Hong[3], Mei Zeng[4]

[1] Nanjing University of Aeronautics and Astronautics, Nanjing, China
[2] Dalian University of Technology, Dalian, China
[3] Beijing Wodong Tianjun Information Technology Co Ltd, Beijing, China
[4] QiLu Hospital of ShanDong University, Jinan, China

Abstract

Sepsis is a life-threatening disease with high mortality and expensive cost of treatment. In order to improve the outcomes of patients, it is important to detect at-risk patients with sepsis at an early stage. The PhysioNet/Computing in Cardiology Challenge 2019 focused on improving predicting sepsis six hours before the clinical diagnosis by using the latest definition of Sepsis-3. A total of 40,336 ICU patients were provided as public training data, A hidden test dataset was used to evaluate. An ensemble model, which combined boosting and bagging tree models (lightgbm, xgboost and random forest) were designed to predict sepsis based on the records of the patient's hourly data. We compared the ensemble model and each single model of evaluation metrics results on selected inner test data Offline, the best performance was achieved AUC of 0.792, ACC of 0.727. Finally, the proposed model was evaluated on the full test sets received an official utility score, defined by the organizers, was 0.087, ranked 75/105 (our team name: cinc_sepsis_pass). While the single model of lightgbm only received a utility score of -0.036. The ensemble model utilized the preprocessing data and achieved better performance than a single tree-based model.

1. Introduction

Sepsis is a life-threatening disease when the body's response to infections cause tissue damage, organ failure or death occurred [1]. Sepsis has become a global public health problem due to high morbidity, mortality and complex pathogenesis, especially in the intensive care unit (ICU). Sepsis is also regarded as a costly disease, the United States cost from $20 billion in 2011 increased to more than $23 billion in 2013, which approximately accounted for 6.2% of all US hospital fee[2]. Early detection and targeted treatment such as antibiotics have been shown are critical to improve sepsis outcomes. Delayed treatment per hour is associated with an approximately 4-8% increase in mortality[3, 4]. The definition of sepsis is also constantly updated. The new recent definition of sepsis-3 is different from the previous criterion, diagnose patients septic if their Sequential Organ Failure Assessment(SOFA) score identified by a two-point deterioration within a 24-hour period[1]. The Physionet/CinC 2019 challenge[5] aims to predict sepsis six hours in advance based on the clinical data.

Most of researches about sepsis focused on the specific patient cohorts and used a different sepsis definition. Calvert et al. proposed a model called InSight for early detection of sepsis by systematic inflammatory response syndrome(SIRs) criteria[6]. Jin's research focused on trauma sepsis patients[7] and Calvert et al.studied high-risk group aged 45 years or older patients for diagnosis of sepsis[8]. The majority of previous work concentrated on a single-center hospital, data mainly from the public MIMIC database[9]. A robust model should be performed similarly when generalized to other hospital systems. So multi-center clinical data provides the possibility of testing for the versatility of the model. Recently many studies have used new definitions of sepsis-3. Nemati [10] demonstrated an interpretable machine learning for predicting sepsis onset 4-12 hour prior to clinical diagnosis. Multi-scale blood pressure and heart rate dynamic feature extraction and Elastic Net logistic model were used to predict sepsis 4 hours prior to its onset by Shashikumar[11], Roman Z Wang [12]compared three models (LR/SVM/LMT) by extracting a random time window 48 to 6 hours prior to the onset of sepsis . Their model mostly applied a single machine learning method,ensemble model is not utilized for early sepsis issue. In addition, evaluation matrices used in these studies were traditional scoring function, such as AUC and AUPRC, those evaluations are not a clinically significant way to reward or punish early detection of false positives or over-treatment. Therefore, the challenge describes a novel evaluate function-utility score [5]to

ISSN: 2325-887X DOI: 10.22489/CinC.2019.317

the problem.The ideal value for the utility score is 1, and higher values indicate better discrimination.

2. Methods

Data was prepared by the organizers of the challenge, which were from three different electronic medical record systems and hospitals. Hospital-A included 20336 patients (sepsis patients: 1790(8.8%), non - sepsis:18546(91.2%)). Hospital-B included 20000 patients (sepsis:1142(5.7%),non-sepsis:18860(94.3%)).The two labeled sets were posted for public download and 24819 patients from three hospital system were sequestered as hidden test sets. Every patient has a file hourly record the clinical data with 40 variables (e.g. heart rate, systolic blood pressure) and 1 sepsis label. (0 means not sepsis in the next 6 hour,1 means sepsis occur in the next 6 hour). A large number of values were missing because measurements were not so frequent and were condensed into hourly bins. Positive and negative samples are extremely imbalance. To build a robust model, it is necessary to make data preprocessing.

2.1. Preprocessing

To get close to the real world of true clinical data, missing and erroneous data were intentionally retained as part of the challenge. Firstly we analyze the distribution of data: The shortest ICU stay record was 8 hours and the longest was 336 hours, most hospital length of ICU stay are 20-35 hours in the two hospitals. In terms of the sepsis patients, we excluded sepsis patients who labeled 1 from the first hour record. 203 and 223 sepsis patients were respectively removed from hospital A and B. Additional sepsis patients whose records with label 1 less than 6 hours were also excluded to prevent a condition that only in the last few hours patients has a label of 1. At last, A/B hospital kept 1587/909 sepsis patients, non-sepsis patients did not change.

Then dealing with the missing data values. We summarized 40 features missing rate and founded variables were missing very badly especially laboratory values. We computed the mean values of each feature from the two hospitals separately. Our impute strategy was using the" pad" method, called carry-forward , where filling the missing value with the previous non-missing value.The overall feature mean values were calculated to fill NaN values. if a patients missing all feature values, then filled with mean value, otherwise used the previous record values to pad. In the medical field, generally speaking, missing value sometimes represents the normal value or keep the same as last measurements.

As for sample imbalances problem, the negative and positive patients ratio was close to 10:1. In our method, every hour record was taken a sample, so we just chose different sample sizes. Not all the negative sample were needed. For example, 1587 sepsis patients record files from hospitals A were used to train, which included 103196 samples (only 15368 hours labeled as 1). To some extent, maintain the same independent distribution as the test set.In this case, According to the score function,it rewards 6-12 hours early prediction, We shifted the label, the first time labeled 1 has been moved forward for 12 hours based on the original label in order to get more rewards.After shifting, for sepsis patients, we just kept the records from admission till the hour of min(T_{sepsis}+3,last record hour).

2.2. Feature extraction

Based on the medical knowledge, we excluded 6 variables :EtCO2,Unit1,Unit2,Gender,BaseExcess, HospAdmTime, which were not related to sepsis or lack of values in hospital system.34 variables were kept and we want to find the remained features which are important and related to sepsis. Three tree-based models (lightgbm,xgboost and random forest) were used to rank the importance of the features by 5-fold cross-validation. The Figure 1 showed the ranked average feature importance by the three models. We conducted two experiments about different numbers of features. First, we used the original 34 features and to add rich representations from the hours records data, maximum, minimum, mean values, and trend information also been calculated. Features were extended to 170 dimensions. The second experiment was used 15 features, which were selected according to feature importance and reference to related literature[13, 14].The experiments in these articles were based on the selection of few important and easily available features for prediction. Besides, trend information about the difference between the predicted current hour and the previous hour was also added. Ultimately, 30 features were used to train the model.

2.3. Model development

With the development machine learning by data-driven, machine learning approaches about medical data were applied widely. An ensemble learning technique that combines multiple base models with a different weights. Different from the homogenous ensembles like bagging and boosting models which use the same type of learner,Sometimes combine and weighted heterogeneous models that can provide better prediction results than a single model. We proposed a ensemble model which combined three different predictive models (lightgbm(lgb),xgboost(xgb) and random forest(rf)) as a baselearner and weighted their output probabilities. The structure of the overall model development was shown in Fig 2.

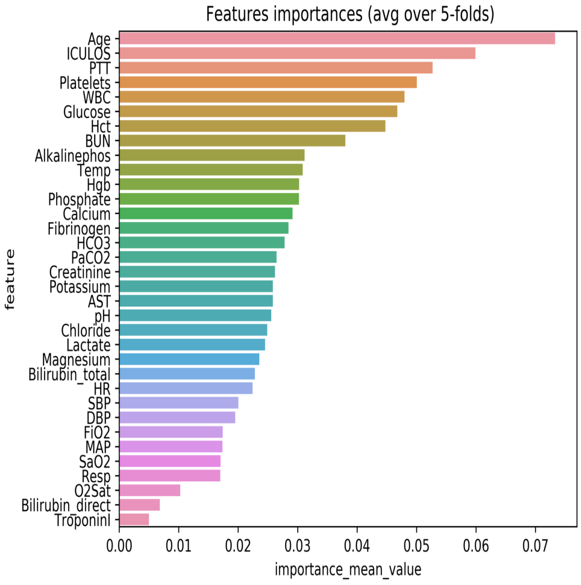

Figure 1. Feature importances score

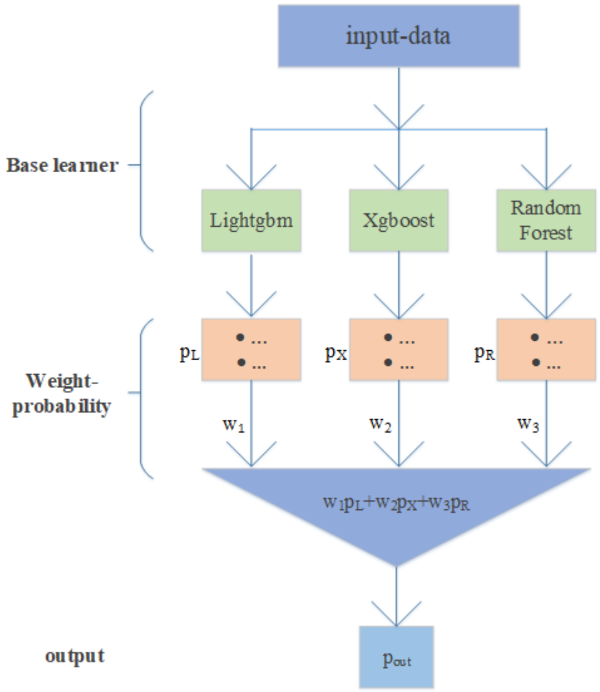

Figure 2. structure of ensemble model. Base-learner included (lgb,xgb,rf) three models,each model trained input data and generated predicted probability then weighted models fusion and got the last output.

3. Results

Limited on the number of submissions, We cant test and compare every single model and the ensemble model with different features and parameters setting online.So we conducted some offline test just used the public download two datasets.

A.online submission(5 times):

	1	2	3	4	5
score on test A	0.018	0.019	0.02	0.04	0.14

Table 1. Submission of five results on leaderboard.

First submission used 170 feature with lightgbm model.The others use selected 30 features, respectively use lightgbm,xgboost,random forest and ensemble model; and the last submission (the sixth time-ensemble model) received the official utility scores on full test was 0.087,on test A/B/C utility scores were 0.154,0.072,-0.155 and AUC values were 0.689,0.719,0.707 respectively;

B.offline experiments:

we conducted three different experiments on A/B sets. use 30 features-15 original features and add trends features, threshold score was set 0.45;

1)Trained on A sets,Test on B set .chosen 1587 sepsis patients files,4000 non sepsis file as training data.total 250425 hourly sample,included 24262 positive samples.Test 2000 files from B,which contained 200 patients,1800 non-patients.

2) Trained on B set,Test on A set.909 patients file,4000 non sepsis files from B sets,about 198472 records,included 12839 positive samples.

3)Mixed A/B sets data.trained 5000 files,selected 1200 sepsis and 2000 non sepsis files from A,600/1200 positive and negative from B sets;21131 entities and 6935 positives hourly data.

	A_train/B_test		B_train/A_test		AB_train/AB_test	
model	AUC	Utility	AUC	Utility	AUC	Utility
lgb	0.751	0.329	0.580	0.243	0.773	0.409
xgb	0.706	0.059	0.591	0.070	0.784	0.140
rf	0.689	0.164	0.581	0.243	0.719	0.255
ensemble	0.744	0.303	0.602	0.259	0.792	0.558

Table 2. offline local train/test result

4. Discussion and Conclusions

Early sepsis prediction for patients in ICU is still a challenging but significant problem. We have developed a ensemble model for the early six hours detection of sepsis from clinical data. From the results of offline test, it was clear that trained mix A/B patients got best performance, the ensemble model achieved a utility score of 0.558 and

AUC value was 0.792. Another interesting conclusion was that the results of models trained with A datasets are generally better than training with B datasets.(e.g. AUC :0.744 versus 0.602,utility score :0.303 versus 0.259). Compared to a single tree-based model, The proposed method provides a new forecasting idea and has a slight improvement utility-based score.Although the prediction effect was not outstanding on the online hidden test datasets. It proved that the ensemble model was better than a single model to a certain extent and provided a new idea to predict sepsis at least. And we also scored the importance of each feature to find the impact factors that are closely related to sepsis. The limitation of the model about generalization ability needs to be improved. Our model could achieve a higher utility score on local offline test,but a worse performance on the online full test sets,which indicated that there was a problem with overfitting. Besides,the parameters and structure of the model also need to be optimized to get a better prediction.

When dealing with real-world clinical data, data preprocessing and feature engineering are greatly important. Not only domain knowledge to build meaningful features are needed, but how to deal with missing and imbalances medical data problem is worth studying. Further studies will be conducted to more exploration and data analysis work.Trying different fill strategies(e.g. K-means clustering, Expectation maximization and Multiple Imputation) for missing values and using oversampling or undersampling methods for unbalanced data processing, those are very worthy of comparison and exploration. Moreover, model fusion methods about integrating deep learning techniques, such as LSTM, which could learn the intrinsic link between time series data, or other machine learning models maybe produce a potential performance improvement.

Acknowledgments

We gratefully acknowledge the assitance of Jiabin Yuan,Peifei Hong,Menglin Lu,Zhenhua Xue in the preparation of this chanllenge and thanks Dr. Zengs medical suggestions.

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] Torio C, Andrews R. National inpatient hospital costs: the most expensive conditions by payer, 2011: statistical brief# 160 2006;.

[3] Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, Suppes R, Feinstein D, Zanotti S, Taiberg L, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Critical care medicine 2006; 34(6):1589–1596.

[4] Seymour CW, Gesten F, Prescott HC, Friedrich ME, Iwashyna TJ, Phillips GS, Lemeshow S, Osborn T, Terry KM, Levy MM. Time to treatment and mortality during mandated emergency care for sepsis. New England Journal of Medicine 2017;376(23):2235–2244.

[5] 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 2019;In press.

[6] Calvert JS, Price DA, Chettipally UK, Barton CW, Feldman MD, Hoffman JL, Jay M, Das R. A computational approach to early sepsis detection. Computers in biology and medicine 2016;74:69–73.

[7] Jin H, Liu Z, Xiao Y, Fan X, Yan J, Liang H. Prediction of sepsis in trauma patients. Burns trauma 2014;2(3):106.

[8] Calvert J, Saber N, Hoffman J, Das R. Machine-learning-based laboratory developed test for the diagnosis of sepsis in high-risk patients. Diagnostics 2019;9(1):20.

[9] Johnson AE, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. Mimic-iii, a freely accessible critical care database. Scientific data 2016;3:160035.

[10] Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the icu. Critical care medicine 2018;46(4):547–553.

[11] Shashikumar SP, Stanley MD, Sadiq I, Li Q, Holder A, Clifford GD, Nemati S. Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics. Journal of electrocardiology 2017;50(6):739–743.

[12] Wang RZ, Sun CH, Schroeder PH, Ameko MK, Moore CC, Barnes LE. Predictive models of sepsis in adult icu patients. In 2018 IEEE International Conference on Healthcare Informatics (ICHI). IEEE, 2018; 390–391.

[13] van Wyk F, Khojandi A, Mohammed A, Begoli E, Davis RL, Kamaleswaran R. A minimal set of physiomarkers in continuous high frequency data streams predict adult sepsis onset earlier. International journal of medical informatics 2019;122:55–62.

[14] Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, Shimabukuro D, Chettipally U, Feldman MD, Barton C, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. JMIR medical informatics 2016;4(3):e28.

Address for correspondence:

Mengsha Fu
School of Computer Science and Technology
No.29 JiangJun Road,Jiangning district, Nanjing, China
mengshafu@nuaa.edu.cn