Early Sepsis Prediction Using Ensemble Learning with Features Extracted from LSTM Recurrent Neural Network

Zhengling He[1,2], Xianxiang Chen[2], Zhen Fang[2], Weidong Yi[1], Chenshuo Wang[1,2], Li Jiang[1,2], Zhongkai Tong[2], Zhongrui Bai[1,2], Yueqi Li[1,2], Yichen Pan[1,2]

[1] University of Chinese Academy of Sciences, Beijing, China
[2] Institute of Electronics, Chinese Academy of Sciences, Beijing, China

Abstract

Early prediction of sepsis can help to identify potential risks in time and help take necessary measures to prevent more dangerous situations from occurring. In PhysioNet/Computing in Cardiology Challenge 2019, we integrate Long Short Term Memory (LSTM) recurrent neural network and ensemble learning to achieve early sepsis prediction. Specifically, we tackle the problem of class imbalance and data missing firstly, and then we manually extract features according to the prior knowledge from the medical field. In addition, we regard the prediction of sepsis as a time series prediction problem and pre-train LSTM-based models as feature extractors to obtain the "deep" features on time series that might be related to the onset of sepsis. Manual features and "deep" features are then used to train prediction models under the framework of ensemble learning, including Extreme Gradient Boosting (XGBoost) and Gradient Boosting Decision Tree (GBDT) regressor. The final normalized utility score our team (UCAS_DataMiner) have obtained was 0.313 on full hidden test set.

1. Introduction

Sepsis and septic shock are important clinical problems among critical patients, its pathogenesis is complex and usually involves infection, inflammation, immune dysfunction. Sepsis can further develop into severe sepsis and septic shock, and the extremely high morbidity and mortality rate have become a serious medical burden. Systemic inflammatory response syndrome (SIRS) criteria has been proposed for the diagnosis of sepsis. However, the SIRS criteria were proved to be inadequate in specificity and sensitivity. So the definitions for sepsis and septic shock were revised based on the epidemiology conducted by the Society of Critical Care Medicine (SCCM) and the European Society of Intensive Care Medicine (ESICM) in 2016. Sepsis was then defined as life-threatening organ dysfunction caused by a host response to infection according to this criteria (sepsis 3.0) [1].

Sepsis causes a large mortality rate in infected patients, especially those who are not recognized and treated promptly. Since the worsening of sepsis is rapid, early prediction can help doctors to take proper treatment to control it, prevent complications and reduce the malignant consequences of shock and even death that may result from the further development of sepsis. In recent years, researches on sepsis have continued to deepen and enhance, and clinical practices and evidences have increased. Electronic health records (EHRs) and machine learning combined approaches are valid ways to achieve early sepsis prediction, support vector machine [2], Bayesian network [3] and some other algorithms have been proposed to early predict the sepsis and help doctors to provide effective medical care and intervention.

The Sequential Organ Failure Assessment (SOFA) score is a widely used method for assessing human or even porcine organ failure [4]. It has also received attention in the research of in-hospital mortality， the relationship between in-hospital mortality and SOFA has been studied and high correlation was reported in some literatures [5, 6]. SOFA scoring system is a new approach for organ failure evaluation and sepsis diagnosis, it is also a critical indicator in sepsis 3.0, a higher SOFA score is associated with an increased probability of mortality, score of 2 points or more means the patient's mortality rate can reach 10%. If the septic shock is further developed, the mortality rate can reach 40%. For patients who have recovered, there is still a high probability of suffering from sequelae.

The goal of PhysioNet/Computing in Cardiology Challenge 2019 is to achieve early prediction of sepsis using multiple clinical parameters, result will be evaluated by the utility score defined by organizers [7]. We proposed an algorithm to achieve the target in this paper, we firstly pre-process the data and manually extract the SOFA and SIRS related features according to the prior knowledge from the medical field. In addition, we construct a structure based on LSTM recurrent neural network, a deep learning method, to automatically learn potential features based on training dataset to mine information on time series. Finally,

 ISSN: 2325-887X DOI: 10.22489/CinC.2019.269

several ensemble learning algorithms, including Extreme Gradient Boosting (XGBoost) and Gradient Boosting Decision Tree (GBDT) models [8, 9], are established as basic learners to gain various predictions, which will be integrated to enhance the accuracy and reduce the bias of final prediction result.

2. Methods

2.1. Data preprocessing

The dataset contains 40 dimensions of original data (F_{RAW}) collected per hour from a total of 40,336 subjects, these records are labeled with 0 or 1 for supervised learning, and the sepsis labels of the data have been shifted ahead by six hours, so we don't need to make any adjustments to the labels.

2.1.1. Data imputation

There are a large number of missing values in the original data, usually due to the absence of related records. If we use the overall mean to replace them, the information related to the individuals can't be revealed. However, it is not feasible to use the individual's mean to fill, due to the fact that we can't get the data after time t in practical situations. So the missing values are filled with its prior adjacent value in the proposed algorithm.

2.1.2. Data imbalance

Data imbalance is another issue that needs to be considered. According to the statistical results, the ratio between labels 0 and 1 exceeds 50, which means serious data imbalance. In order to reduce the impact of data imbalance on the results, we discard the data with the label 0 with a probability of 1/5 in the data reading phase. The data will be further randomly down sampled again while training, so that the 0 and 1 classes can keep balance. On the other hand, the data perturbation enhances the diversity of the model and contributes to the subsequent ensemble learning.

2.2. Features extraction

2.2.1. Features based on definitions of sepsis

The latest sepsis 3.0 criteria defines the qSOFA and SOFA scores. qSOFA criteria is simple (Respiratory rate>=22/min and SBP<=100 mmHg), so it provides a quick diagnostic method for doctors and can help promote diagnostic efficiency. We convert the rules defined above into binariztion features, that is, when a parameter (such as blood pressure, respiration rate, etc.) exceeds the defined threshold, this feature will be recorded as 1, otherwise it is recorded as 0. We also manually calculate the SOFA score according to the sepsis 3.0 criteria [1]. It should be noted that some parameters are missing, such as Fraction of inspired oxygen (FIO2), so the SOFA we obtain is actually a pseudo-SOFA, however, intuitively we believe that higher pseudo-SOFAs are still associated with sepsis prediction. In order to further obtain the trend of all parameters defined in the criteria throughout the measurement process, the ratio between the current value v_t at time t and the baseline value v_0 of each individual is calculated, considering the missing of some values, the baseline value is firstly initialized to *-99*, and the ratio will be set as follows:

$$ratio_t = \begin{cases} 0, & v_0 == \text{-}99 \\ (v_t - v_0)/v_0, & v_0 != \text{-}99 \ and \ v_t != Null \\ ratio_{t-1}, & v_0 != \text{-}99 \ and \ v_t == Null \end{cases} \quad (1)$$

All the above features obtained are denoted as F_{SOFA} .

SIRS is another criteria that has been used in the detection of sepsis, although it has been replaced by the latest criteria, it may still remain useful for the identification of infection. So we convert the rules according to its definition into binariztion features, and the ratio between v_t and v_0 is also calculated, and these features are denoted as F_{SIRS} .

2.2.2. LSTM-based features

RNN based models have achieved great success in numerous fields especially time series related problems, the foremost advantage is that it can easily extract the dependency information on the time series. Long Short Term Memory (LSTM) is a specially designed recurrent neural network, which is mainly used to solve the problem of vanishing gradient of traditional recurrent neural networks that limit long-term dependence and has been widely applied in numerous applications [10-12]. The reason we use LSTM is that the onset of sepsis is a sequential process, and thus it should be not only related to the features extracted from the current time, but also should be related to the information extracted from the previous hours.

In detail, we firstly stack multiple LSTM layers with fully connected (FC) layers together to form a deep structure with powerful skill of feature representation, and then we pre-train this model on a subset of training dataset, the training process is monitored manually to prevent both under-fitting and over-fitting. Finally, we remove the output layer and only retain the last hidden layer as the feature extractor to extract LSTM-based features.

2.3. Model architecture

The structure of the proposed algorithm is shown in Figure 1. Data are randomly divided by the ratio of 9:1 for

training and testing. LSTM-based models are firstly trained on a subset of 10,000 subjects from training dataset. For the remaining subjects, the data will be forwarded to the pre-trained LSTM extractor and gain output F_{LSTM}.

In order to enhance the variability of the models, we construct various features subsets. Specifically, not all features are entered into the level-2 regressor, but a subset of them are selected and concatenated together as the input vector to the regressor. The final feature subsets including: F_{RAW}, $F_{RAW}+F_{LSTM}$, $F_{RAW}+F_{SOFA}$, $F_{RAW}+F_{SIRS}$, $F_{RAW}+F_{SIRS}+F_{SOFA}+F_{LSTM}$. As for the level-2 regressor, we tried support vector regression (SVR), linear regression, etc. Finally, the XGBoost and GBDT models outperformed than others, so we trained 10 models based on these two architectures and ensemble the results to get the final probability as output.

The LSTM-based features extractors are implemented by Keras with TensorFlow backbend and level-2 regressors are implemented by Scikit-Learn package.

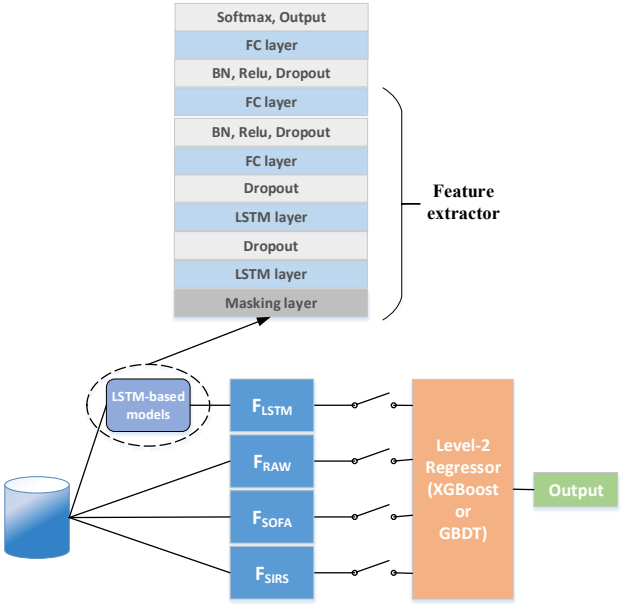

Figure 1. The whole structure of the algorithm.

2.3.1. LSTM-based features extractor

Each LSTM-based model consists of P LSTM layers and Q fully connected (FC) layers with Batch-normalized (BN) and Dropout layers embedding. RELU is selected as activate function except the output layer, and the last layer use Softmax activation. Data will be firstly converted to a tensor like $[sample, timestep, \dim ension]$. However, when $t < timestep$, we can't construct the input tensor with $timestep$, Keras provides effective approach, which is called "Masking" layer, to tackle this problem. So we add a "Masking" layer ahead of the input layer to adapt to time series with different time steps.

We enhance the diversity of LSTM extractors by adjusting the hyperparameters of the model, for example, using different *timestep* (The optimal *timestep* we chose were 4 to 6 hours), adjusting the number of neurons in each hidden layer or the number of hidden layers, and finally we obtained N_{LSTM} basic "deep" features extractors.

We use binary cross-entropy loss as the optimization target, and Adam is applied to the back propagation process with initial learn rate of 0.001 to optimize it, and Early-Stopping criteria is also used to prevent over-fitting.

2.3.2. Ensemble learning

Various ensemble learning models have been proposed and applied to different applications [13-15]. The advantage of the ensemble learning is the ability to combine multiple basic models to achieve an integrated model for more accurate, stable and robust results.

Gradient Boosting framework boosts numerous weak prediction models (such as linear model, decision tree) to a more powerful one. XGBoost and GBDT are gradient boosting machines based on tree model as the basic weak predictor, they have been widely used in various classification and regression tasks. Our experimental results show that they also perform well in the task of sepsis prediction. We use grid search to find the optimal hyperparameters.

From the perspective of the overall algorithm, we can further perform "ensemble learning". In detail, we regard the level-2 regressor as a strong model (low deviation but high variance) and make various predictions firstly by pre-training multiple models with different structure, and then we integrate them by simple averaging to obtain the final results, as shown in Eq. 2.

$$EnsembleProb = \sum_{i=1}^{N} \frac{BaseModel\_Prob_i}{N} \quad (2)$$

Where $BaseModel\_Prob_i$ is the prediction probability obtained by *i-th* model, N is the number of basic models.

3. Results and discussion

We verified and adjusted our algorithm on officially published datasets A and B (including 40,336 subjects) by 5-fold cross validation, and the final normalized utility score our team (UCAS_DataMiner) have obtained is 0.313 on full hidden test set (0.406, 0.373 and -0.215 on test set A, B and C, separately). The cross validated averaged utility score on XGBoost and GBDT regressor with different features subsets were used to evaluate the performance of each model, as shown in Table 1.

It is notable that different feature subsets and models do not differ significantly in averaged utility score, yet they provide different perspectives for the final ensemble model, in other words, establish multiple feature subspaces that

are beneficial for ensemble learning, so the averaged utility score has been improved to 0.401 in the ensemble model. Actually, according to error-ambiguity decomposition [16], the higher the accuracy and diversity of the basic model is, the better the ensemble results. So the measures we mentioned above, including randomly dividing the data for training and testing, constructing 5 different feature subsets, and increasing the diversity of the LSTM extractor, are all for this purpose.

Table 1. 5-fold cross validated averaged utility score on XGBoost and GBDT regressor with different features subsets.

Features Subsets	Averaged utility score	
	XGBoost	GBDT
F_{RAW}	0.387	0.385
F_{RAW}+F_{LSTM}	0.385	0.386
F_{RAW}+F_{SOFA}	0.389	0.379
F_{RAW}+F_{SIRS}	0.390	0.384
F_{RAW}+F_{LSTM}+F_{SOFA}+F_{SIRS}	0.380	0.382
Ensemble	**0.401**	

4. Conclusion

In this paper, we propose a novel algorithm under the framework of ensemble learning to integrate manual features and LSTM-based "deep" features to achieve early sepsis prediction. The current results on hidden test dataset demonstrate the validity of the model, and we have obtained a normalized utility score of 0.313 on full hidden test set. Future work will focus on integrating more features from the medical field and experimenting with more machine learning methods to further improve the accuracy, robustness of prediction.

References

[1] M. Singer *et al.*, "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," *JAMA,* vol. 315, no. 8, pp. 801-10, Feb 23 2016.

[2] S.-L. Wang, F. Wu, and B.-H. Wang, "Prediction of severe sepsis using SVM model," in *Advances in computational biology*: Springer, 2010, pp. 75-81.

[3] L. Peelen, N. F. De Keizer, E. De Jonge, R.-J. Bosman, A. Abu-Hanna, and N. Peek, "Using hierarchical dynamic Bayesian networks to investigate dynamics of organ failure in patients in the Intensive Care Unit," *Journal of biomedical informatics,* vol. 43, no. 2, pp. 273-286, 2010.

[4] K. E. Soerensen *et al.*, "The use of sequential organ failure assessment parameters in an awake porcine model of severe Staphylococcus aureus sepsis," *Apmis,* vol. 120, no. 11, pp. 909-921, 2012.

[5] D. A. Geerse, L. F. Span, S.-j. Pinto-Sietsma, and W. N. van Mook, "Prognosis of patients with haematological malignancies admitted to the intensive care unit: Sequential Organ Failure Assessment (SOFA) trend is a powerful predictor of mortality," *European journal of internal medicine,* vol. 22, no. 1, pp. 57-61, 2011.

[6] A. E. Jones, S. Trzeciak, and J. A. Kline, "The Sequential Organ Failure Assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation," *Critical care medicine,* vol. 37, no. 5, p. 1649, 2009.

[7] Reyna MA, Josef C, Jeter R, Shashikumar SP, 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.

[8] F. J. H. . "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics,* vol. 29, no. 5, pp. 1189-1232, 2001.

[9] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," pp. 785-794, 2016.

[10] S. Chauhan and L. Vig, "Anomaly detection in ECG time signals via deep long short-term memory networks," in *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2015, pp. 1-7: IEEE.

[11] S. L. Oh, E. Y. Ng, R. San Tan, and U. R. Acharya, "Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats," *Computers in biology and medicine,* vol. 102, pp. 278-287, 2018.

[12] M. Cheng, W. J. Sori, F. Jiang, A. Khan, and S. Liu, "Recurrent neural network based classification of ecg signal features for obstruction of sleep apnea detection," in *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, 2017, vol. 2, pp. 199-202: IEEE.

[13] G. Wang, J. Hao, J. Ma, and H. Jiang, "A comparative assessment of ensemble learning for credit scoring," *Expert systems with applications,* vol. 38, no. 1, pp. 223-230, 2011.

[14] L. Yu, S. Wang, and K. K. Lai, "Credit risk assessment with a multistage neural network ensemble learning approach," *Expert systems with applications,* vol. 34, no. 2, pp. 1434-1444, 2008.

[15] J. Miskin and D. J. MacKay, "Ensemble learning for blind image separation and deconvolution," in *Advances in independent component analysis*: Springer, 2000, pp. 123-141.

[16] F. Schwenker, "Ensemble Methods: Foundations and Algorithms [Book Review]," *IEEE Computational Intelligence Magazine,* vol. 8, no. 1, pp. 77-79, 2013.

Address for correspondence:

Xianxiang Chen, Zhen Fang
Institute of Electronics, Chinese Academy of Sciences, No. 9, North Zhongguancun, Haidian District, Beijing.
xxchen@mail.ie.ac.cn, zfang@mail.ie.ac.cn