Representation Learning for Early Sepsis Prediction

Luan Tran[1], Manh Nguyen[2], Cyrus Shahabi[1]

[1] University of Southern California, Los Angeles, USA
[2] 3S Intersoft JSC, Hanoi, Vietnam

Abstract

As part of the PhysioNet/Computing in Cardiology Challenge 2019, we propose a neural network called AEC-Net to early detect sepsis based on physiological data. AEC-Net consists of two main components: 1) an Auto Encoder for dimension reduction and feature extraction, and 2) a Fully Connected Neural Network (FCNN) taking the extracted features by the Auto Encoder as the input and generating prediction of sepsis as output. The losses of both the Auto Encoder and FCNN are minimized concurrently. This concurrent optimization helps AEC-Net to have a better generalization and the extracted features by Auto Encoder to be more relevant to the classification problem. Finally, we propose an ensemble method of AEC-Net, Random Forest and Gradient Boosting Decision Trees to achieve a better prediction.

We evaluate our proposed methods using data from 40336 patients with 40 physiological features ranging from 8 to 336 hours. AEC-Net outperforms RF and GBDT in Utility, a metrics defined by the Physionet Challenge 2019, of 0.39. Our proposed ensemble algorithm achieves an AUC_ROC of 0.82, Accuracy of 0.90, F-measure of 0.16, AU_PRC of 0.08, and Utility of 0.41. It outperforms all the single classifiers in every metrics. With the hidden test data of the Physionet Challenge 2019, Ensemble achieved a Utility of 0.38.

1. Introduction

Sepsis is a serious condition resulting from the presence of harmful microorganisms in the blood potentially leading to the malfunctioning of various organs, shock, and death. With every hour of delay treatment after the onset of hypo-tension, the risk of mortality from septic shock increases by 7.6 % [1]. Therefore, an early intervention is crucial. Since the electronic health records are widely adopted, there exists a wealth of data to inform predictions about when sepsis is likely to occur. However, early sepsis prediction is still challenging because the symptoms in the physiological data associated with sepsis can be also caused by many other clinical conditions. In this study, we focus on predicting sepsis conditions 6 hours earlier than the clinical detection using physiological data such as temperature, heart rate, and FiO_2. More specifically, at hour t, the problem is to predict whether a patient will have sepsis at hour $t + 6$ using data from the past until hour t. This problem can be formalized as a binary classification problem of multi-dimensional time series. With the advance of machine learning, especially deep learning, many classification methods have been introduced such as Random Forest (RF) [2], Gradient Boosting Decision Trees (GBDT) [3], Deep Neural Network (DNN) [4] for structured data, and LSTM network [5] for time series data. The traditional machine learning algorithms (e.g., RF, GBDT and DNN) need feature engineering which requires domain knowledge. Auto Encoder, a deep neural network, was introduced as an efficient method for automatic feature extraction. Therefore, in this study, we design a neural network which consists of an Auto Encoder and a Fully Connected Neural Network (FCNN). FCNN takes the extracted features by the Auto Encoder as the input and outputs the class label, healthy or has-Sepsis. Since we want the extracted features to be relevant to the classification, we optimize the Auto Encoder concurrently with FCNN by combining their two losses. The contributions of our study are summarized as follows:

1) We implement two machine learning classifiers: Gradient Boosting Decision Trees and Random Forest, as baselines.

2) We propose a deep neural network, AEC-Net, which optimizes concurrently an Auto Encoder and a Fully Connected Neural Network. We show that AEC-Net outperforms the baseline algorithms.

3) We propose an ensemble framework of AEC-Net, RF, and GBDT to take advantage of all the classifiers. This ensemble achieves higher performance than any single classifier in all the metrics, i.e., AUC_PRC, AU_PRC, Accuracy, F-measure and Utility [1].

The remainder of the paper is organized as follows. In Section 2, we present a literature review. In Section 3, we discuss data imputation and feature extraction. In Sections 4 and 5, we review our implementation of Gradient Boosting Decision Trees and Random Forest, respectively. In Section 6, we introduce our proposed deep neural network. In Section 7, we explain our ensemble method. In Section 8, we report our experiments and the results. Finally, in

[1] https://physionet.org/challenge/2019/

Section 9, we conclude the paper.

2. Literature Review

To detect sepsis, researchers have introduced several scoring systems such as SIRS [6], MEWS [7], SOFA [8], and QSOFA [9]. They are based on some criteria on physiological data such as heart rate is higher than 90 beats/min and body temperature is higher than 38^oC degree. Recently, several machine learning models such as InSight [10], SVM [11], and Deep Neural Network [12], are used in early prediction of sepsis, and they are shown to be better than the traditional scoring systems. All of the previous studies try to predict sepsis 0-4 hours earlier than the clinical detection.

3. Data Imputation and Feature Extraction

In this study, we use the data from the Physionet 2019 Challenge [2] with the patients' records of 40 features in multiple hours. The percentage of missing values for each column is reported in Table 1. We replace the missing values in each column with the mean value of that column for the entire dataset. After imputing missing values, we use

Missing Value Percentage	>25%	>50%	>75%
Number of Columns	31	28	27

Table 1. Data Missing Percentage

the data at hour $t-5$ to t to classify the data at hour t. We will observe the groundtruth at time $t+6$. Since there are 40 features, for the AEC-Net, the input format is $(6, 40)$. And for GBDT, Random Forest, we flatten the input data into the format of $(1, 240)$.

Number of Patients without Sepsis	37404
Number of Patients with Sepsis	2932
Total number of Patients	40336

Table 2. Patient data statistics

Table 2 depicts the number of patients with and without sepsis in the dataset. This illustrates the imbalance in the dataset. To resolve the problem of class imbalance, we use all data with sepsis and only 5% of data without sepsis for training.

4. Gradient Boosting Decision Trees

Gradient Boosting Decision Trees (GBDT) [3] is a method of converting weak learners (decision trees) into strong learners. Starting from a decision tree, this method subsequently adds new decision trees to have a stronger classifier. Specifically, a decision tree outputs real values for splitting at the internal nodes of the tree and their outputs can be added together. This allows for adding subsequent decision trees based on the residuals of the predictions of the current trees. When adding a new decision tree, GBDT performs the gradient descent procedure on its parameters to reduce the total loss.

5. Random Forest

Random Forest (RF) [2] is a model made up of several decision trees. This method averages the predictions of a number of decision trees. During the training phase, each decision tree is built with a random subset of features. The number of features considered in each subset of features is a tunable parameter. Besides, there are several other parameters which are tunable, such as the number of estimators and the maximum depth of tree.

6. Deep Neural Network - AEC-Net

In this section, we present our proposed deep neural network, AEC-Net. The structure of AEC-Net is depicted in Figure 1. Instead of classifying data in the original space $\mathbf{X}$, we propose to first transform data with a non-linear mapping $f_\theta : \mathbf{X} \rightarrow \mathbf{Z}$ where θ is a list of learnable parameters and $\mathbf{Z}$ is the latent *feature space*. The dimensions of $\mathbf{Z}$ is typically much smaller than $\mathbf{X}$ in order to avoid "the curse of dimensionality" [13]. To parameterize f_θ, Deep Neural Network (DNN) is a natural choice due to its theoretical function approximation property [14] and its demonstrated feature learning capability [15]. Following this approach, our proposed network consists of two main components, i.e., an LSTM Auto Encoder [16] and a Fully Connected Neural Network (FCNN). Auto Encoder, a deep neural network, which has been shown to be efficient in representation learning, is used to extract latent compact features from the input data. The LSTM Auto Encoder consists of an encoder and a decoder. The encoder contains three LSTM layers with 128, 64, and 64 cells, respectively. LSTM [3] is used to capture the temporal dependency in the data. The decoder also contains three LSTM layers with 64, 64, and 128 cells, respectively. The decoder reconstructs the input data from the encoded data by minimizing the mean square error loss:

$$\mathbf{L_1} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{x_i} - \mathbf{\hat{x}_i})^{\mathbf{2}} \tag{1}$$

where $\mathbf{n}$ is the number of training data and $\mathbf{x_i}$ and $\mathbf{\hat{x}_i}$ are the input data and the reconstructed data, respectively. The output of the last layer of the encoder is used as extracted features. FCNN takes the features extracted by the LSTM Auto Encoder as its input and output the class label, healthy or has-Sepsis. Our FCNN consists of 5 fully connected layers with 32, 24, 16, 8 and 2 units, respectively.

[2] https://physionet.org/challenge/2019/

[3] https://keras.io/layers/recurrent/

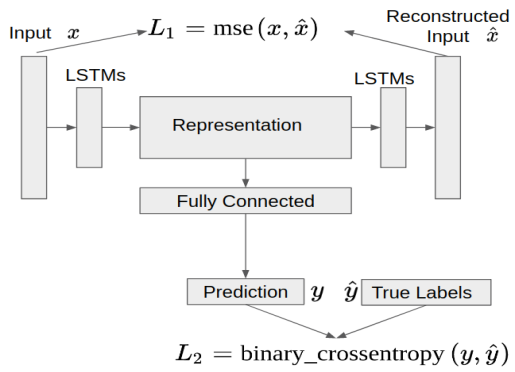

Figure 1. AEC-Net Structure

Before each fully connected layer, we apply a Dropout layer with the dropout rate of 0.3 for reducing over-fitting. FCNN is optimized by minimizing the binary cross entropy loss:

$$\begin{aligned} \mathbf{L_2} &= -\frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{j=1}}^{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{2}} \mathbf{t_i log(s_i)} \\ &= -\frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{j=1}}^{\mathbf{n}} \mathbf{t_1 log(s_1) - (1 - t_1) log(1 - s_1)} \end{aligned} \tag{2}$$

where $\mathbf{n}$ is the number of training data, $\mathbf{t_i}$ and $\mathbf{s_i}$ are the ground-truth and score for class $\mathbf{i}$, respectively.

These two sub-networks are optimized concurrently by linear combination of the two losses, in order to force the LSTM Auto Encoder to learn relevant features to the classification task:

$$\mathbf{L = w_1 \times L_1 + w_2 \times L_2} \tag{3}$$

where $\mathbf{w_1}, \mathbf{w_2}$ are the weights measuring the importance of the tasks and $\mathbf{w_1 + w_2 = 1}$.

7. Classifier Ensemble

Each classifier is able to detect different cases of sepsis. Therefore, we propose to combine the classifiers as an ensemble of AEC-Net, Random Forest, and Gradient Boosting Decision Trees, to achieve a better performance in predicting sepsis. If any classifier outputs have-sepsis, the overall predicted label is have-sepsis, or more formally:

$$\mathbf{Ensemble(x) = AEC(x) \vee RF(x) \vee GBDT(x)} \tag{4}$$

where $\mathbf{x}$ is the input data, $\mathbf{AEC(x)}$, $\mathbf{RF(x)}$, and $\mathbf{GBDT(x)}$ are the output of AEC-Net, RF and GBDT, respectively.

	AEC-Net	RF	GBDT	Ensemble
AUC_ROC	0.81	0.81	0.82	0.82
Accuracy	0.86	0.93	0.91	0.9
F-measure	0.12	0.17	0.16	0.16
AU_PRC	0.11	0.08	0.09	0.08
Utility	**0.39**	0.33	0.35	**0.41**

Table 3. Classifier Comparison

8. Experiment Results

8.1. Experimental Methodology

We randomly separated the dataset into a training and testing dataset. The training dataset contains 27025 samples and the testing dataset contains 13311 samples. We also evaluated our best classifier with the hidden test data provided by the PhysioNet Challenge 2019.

8.2. Parameter Settings

For GDBT and RF, we set the number of estimators to 250. For AEC-Net, we train the model using an Adam optimizer [4] with the learning rate of 0.001 and the batch size of 256. The weights $\mathbf{w_1}$ and $\mathbf{w_2}$ corresponding to the two losses are set to 0.2 and 0.8, respectively, to give a larger weight on the classification task.

8.3. Classifier Comparison

We compare AUC_ROC, Accuracy, F-measure, AU_PRC and Utility [5] of all the methods. The intuition behind the Utility metric is to give higher score for correct sepsis detection and higher penalty for undetected sepsis cases. Table 8.3 shows the AUC_ROC, Accuracy, F-measure, AU_PRC and Utility with the test data for all the classifiers. As illustrated in this table, AEC-Net, RF and GBDT show comparable AUC_ROC, Accuracy, F-measure and AU_PRC. AEC-Net achieves higher Utility than RF and GBDT. Ensemble achives the highest performance in all metrics since it detects more sepsis cases that than those of any single classifier.

Since Ensemble offers the highest Utility, we evaluated Ensemble with the hidden test dataset provided by the Physionet Challenge 2019. We achieved a Utility of 0.38.

8.4. The Impact of Auto Encoder

In this section, we verify the usefulness of Auto Encoder in finding a good representation for the input data. We compared the performance of AEC-Net with and without the decoder. When removing the decoder, the raw input data is passed through LSTM layers of the Encoder and

[4] https://keras.io/optimizers/
[5] https://physionet.org/challenge/2019/

	With Decoder	Without Decoder
AUC_ROC	0.80	0.79
Accuracy	0.87	0.89
F-measure	0.13	0.13
AU_PRC	0.09	0.08
Utility	**0.38**	0.36

Table 4. With and Without Decoder Comparison

the Fully Connected Neural Network. For this network, we only minimize the classification loss. Table 4 shows the results of AEC-Net with and without the decoder sub-network. As depicted in this table, removing the decoder sub-network reduces the performance of the network, e.g., by 0.2 in Utility. Without the decoder sub-network, the network's ability for generalization is reduced.

9. Conclusions

In this paper, we proposed a deep neural network AEC-Net which optimizes an LSTM Auto Encoder and a Fully Connected Neural Network concurrently. Our experimental results with real-world datasets showed that AEC-Net outperforms the two baselines RF and GBDT in terms of Utility. We also proposed an ensemble method consisting of AEC-Net, RF, GBDT and our results showed that it outperforms any single classifier in all the metrics.

10. Acknowledgements

This work has been supported in part by the USC Integrated Media Systems Center, and unrestricted cash gifts from Oracle and Google. The opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1] 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.

[2] Liaw A, Wiener M, et al. Classification and regression by randomforest. R news 2002;2(3):18–22.

[3] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. Lightgbm: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems. 2017; 3146–3154.

[4] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems. 2012; 1097–1105.

[5] Graves A, Mohamed Ar, Hinton G. Speech recognition with deep recurrent neural networks. In 2013 IEEE international conference on acoustics, speech and signal processing. IEEE, 2013; 6645–6649.

[6] Bone RC, Balk RA, Cerra FB, Dellinger RP, Fein AM, Knaus WA, Schein RM, Sibbald WJ. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. Chest 1992;101(6):1644 – 1655. ISSN 0012-3692.

[7] Subbe C, Kruger M, Rutherford P, Gemmel L. Validation of a modified early warning score in medical admissions. Qjm 2001;94(10):521–526.

[8] Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, Reinhart C, Suter P, Thijs L. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. Intensive care medicine 1996; 22(7):707–710.

[9] Barbara P, Graziano C, Caputo W, Litvak I, Battinelli D, Hahn B. The quick sequential organ failure assessment (qsofa) identifies septic patients in the out-of-hospital setting. The American journal of emergency medicine 2018; 36(6):1022–1026.

[10] Mao Q, Jay M, Hoffman JL, Calvert J, Barton C, Shimabukuro D, Shieh L, Chettipally U, Fletcher G, Kerem Y, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and icu. BMJ open 2018;8(1):e017833.

[11] Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. PloS one 2017;12(4):e0174708.

[12] Kam HJ, Kim HY. Learning representations for the early detection of sepsis with deep neural networks. Computers in Biology and Medicine 2017;89:248 – 255. ISSN 0010-4825.

[13] Indyk P, Motwani R. Approximate nearest neighbors: towards removing the curse of dimensionality. In Proceedings of the thirtieth annual ACM symposium on Theory of computing. ACM, 1998; 604–613.

[14] Hornik K. Approximation capabilities of multilayer feedforward networks. Neural networks 1991;4(2):251–257.

[15] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence 2013;35(8):1798–1828.

[16] Marchi E, Vesperini F, Eyben F, Squartini S, Schuller B. A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional lstm neural networks. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015; 1996–2000.

Address for correspondence:

Luan Tran
Computer Science Department
University of Southern California, Los Angeles, CA, USA.
luantran@usc.edu