Novel Imputing Method and Deep Learning Techniques for Early Prediction of Sepsis in Intensive Care Units

E Macias[1], G Boquet[1], J Serrano[1], JL Vicario[1], J Ibeas[2], A Morel[1]

[1] Wireless Information Networking, Universitat Autònoma de Barcelona, Bellaterra, Spain
[2] Nephrology Department, Parc Taulí Hospital Universitari, Institut de Investigació i Innovació Parc Taulí I3PT. Universitat Autònoma de Barcelona, Sabadell, Spain

Abstract

It is possible to exploit the predictive capacity of data collected in intensive care units (ICU) with a high ratio of missing values. Combining several sources of information, a considerable number of missing values are generated. In this manuscript, an alternative approach to impute this type of data, together with the use of deep learning techniques to improve the early detection of sepsis in ICU is proposed. Initially, laboratory tests are separated and summarized. Then, their most representative information is extracted by taking codes from an autoencoder. This information is combined with the rest of the variables and used to exploit temporal dependencies through long short-term memory recurrent neural networks. With the proposed approach our team, WIN-UAB, was ranked in the position 38/78 with a utility score (defined in the the PhysioNet/Computing in Cardiology Challenge 2019) of 0.241 on the full test set. The predictive capacity of the proposed solution demonstrated the potential of integrating an alternative approach for imputing variables with a high ratio of missing values. In terms of dimensionality reduction, it is possible to reduce 27% of features through the codes of autoencoders.

1. Introduction

In the era of machine learning, it is possible to capture and extract knowledge from large amounts of medical data. The evolution of patients composed of different types of registers such as images, vital signs, diagnoses, among others. The follow-up of variables depends on several factors, such as the type of disease and the determinations of clinical samples. In this way, a patient will have a mixture of variables that will rarely be taken at the same time.

A clear case of this issue happens in intensive care units (ICU), where vital signs are monitored continuously, while laboratory tests are taken less frequently. Combining several sources of information along with possible errors in measurement, equipment failure, lack of collections, or determinations that do not match their timestamp, generate a considerable amount of missing values. On the other hand, one of the most critical problems in ICU is sepsis and its challenging early detection [1]. It represents an epidemiologic problem, with more than 30 million people who develop it and more than 6 million who die every year [2]. Although several works have started using machine learning techniques for the detection of pathologies [3–6], the most widely way to identify them is through clinical scores that relate the risk factors with events linearly [7]. However, the applications of other strategies to deal with a high ratio of missing values and more complex models, that take advantage of non-linear relationships can improve the detection of sepsis and be truly useful in the medical domain.

Thus, defining mechanisms that exploit the predictive capacity of the data with high ratios of missing values, together with the temporary dependencies resulting from the monitoring of the patients in ICU, it is possible to improve the early detection of pathologies to support the clinical decisions. From massive data of patients and their progression in ICU, in this manuscript, it is proposed to combine a novel mechanism to impute variables with high ratios of missing values and use deep learning (DL) techniques for the early detection of sepsis in ICU. In summary, the main contributions of this manuscript are:

- Impute in novel way variables with a high ratio of missing values.
- Extract the most relevant information from the data and reduce its dimensionality through autoencoders.
- Exploit the predictive capacity of temporal evolution using long short-term memory (LSTM) recurrent neural networks (RNN).

2. Materials and methods

This work is carried out with a cohort of 40336 patients admitted to two ICUs from the United States so-called A

ISSN: 2325-887X DOI: 10.22489/CinC.2019.038

and B. To avoid bias in the results, initially only part of the A and B data was made public by the organizers of the *Early Prediction of Sepsis from Clinical Data – the PhysioNet Computing in Cardiology Challenge 2019* [1]. There is a third unit (C) whose data is completely hidden to test the models. Each patient has 41 variables related to demographics (6), laboratory tests (26), vital signs (8), and the target, which refers to the development of sepsis in the ICU. Each register contains one hour of follow-up of the patients. The goal of the challenge was to detect sepsis 6 hours before it occurred. For this, Figure 1 shows the methodology used in this work to solve this problem.

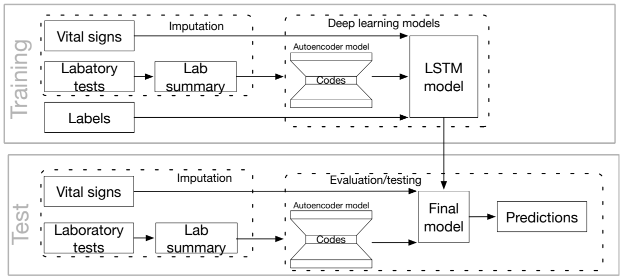

Figure 1. Work flow for the early prediction of sepsis in ICU.

Initially, patients are randomly separated for training and test sets (70-30%). Then, due to the high ratio of missing values, laboratory tests and vital signs are imputed in different ways. Finally, DL models are applied to extract the most relevant information and exploit the temporal dependencies through AE and LSTM, respectively. Next are described in detail the necessary steps to extract and combine information from the available variables using DL techniques to perform the early detection of sepsis.

2.1. Imputation

Due to the different missing values rates in laboratory tests and vital signs (see Figure 2), they are imputed separately. For laboratory tests, a window of $'N'$ hours is taking and then summarized in one register, that is, the imputer value for a variable. In the case of having several values in the window, the imputer shall be the mean of the variable in the window. For vital signs, considering they are monitored continuously, they are imputed using second-order interpolation. In both cases, for patients without determinations, their variables are imputed using the mean value of the variable from the training set.

Once the imputation is done, AE is trained to extract the most important information from the laboratory tests through its codes and finally are merge with the rest of the variables to feed an LSTM model.

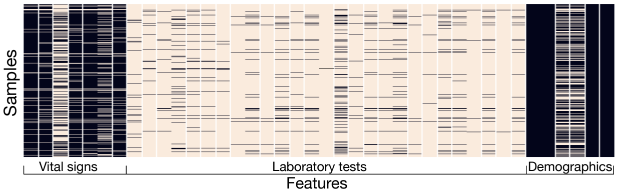

Figure 2. Missing values distribution for all the variables. Dark regions refer to the available data.

2.2. Deep learning models

DL models are based on artificial neural networks (ANN) with more than one hidden layer. Its goal is to learn a non-linear model that maps the input $\mathbf{x_n}$, where $n = 1, ..., N$ to its corresponding targets $\mathbf{t_n}$. The error between predicted output and the target is measured through a cost function. For this work, the mean square error (MSE) for AE and the binary cross-entropy (C($\mathbf{W}$)) for LSTM, Eq 1 and Eq 2 respectively, are used.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{y_i} - \mathbf{y_i'})^{\mathbf{2}} \tag{1}$$

where $\mathbf{y_i}$ is the target and $\mathbf{y_i'}$ are the predicted values.

$$C(\mathbf{W}) = -\sum_{\mathbf{n=1}}^{\mathbf{N}}\sum_{\mathbf{k=1}}^{\mathbf{K}} \mathbf{t_{kn}}\mathbf{log}(\mathbf{y_k}(\mathbf{x_n}, \mathbf{W})) \tag{2}$$

where *N* is the number of samples, *K* is the number of classes, $y_k(\mathbf{x_n}, \mathbf{W})$ is the softmax outputs, and t_{kn} the binary target values.

In both cases, the training is carried out minimizing the cost function iteratively. It is based on forward and back-propagation. Initially, the input data is spread across the network. The weights of each connection are multiplied by their input and added to a bias term. This product called activation, $a_j = \sum_i w_{ji}x_i + b_j$, is then passed through a non-linear function that transforms it to a range of values, typically between [0, 1] or [-1, 1]. The most commonly used activation functions are Sigmoid, hyperbolic tangent (tanh) or rectified linear unit (ReLU). Once these values reach the output layer of the network, it is decided if the error is small enough for training. If not, the weights of the network are updated with the information of the gradient of the cost function; see Eq 3.

$$\mathbf{W(t+1) = W(t) - LR * \Delta C(W(t))} \tag{3}$$

To speed-up the learning process, learning rate (LR), which controls how fast the error is moving to a local minimum, is dynamically changed by optimizers. In this work, adaptative moment estimation (ADAM) is used. It uses first and second-order momentum to update the

LR at each iteration. To avoid overfitting, a common problem on DL models, some techniques such as early stopping, increasing the dataset, or applying regularizers are applied. In this work, L2 regularization is used to penalize the weights that tend to be very large, which avoid the generalization of the learning model.

That said, we make use of two different DL models. The AE to extract the essential information of laboratory tests in a smaller space and the LSTM-RNN to exploit the temporal evolution of patients in the ICU.

2.2.1. Autoencoders

AE are a type of ANN that works in an unsupervised way. Its goal is to replicate the input **x** to the output, **x'**, with the smallest error. The input is forced to go through layers with fewer dimensions, and then the AE has to reconstruct it. As the MSE error is minimized, the output of the hidden layers contains essential information with fewer dimension. The network is composed of two parts, an encoder function **h**=f(**x**) and a decoder that produces the reconstruction **x'**=g(**h**). The encoder function generates the so-called codes, that best represents the data.

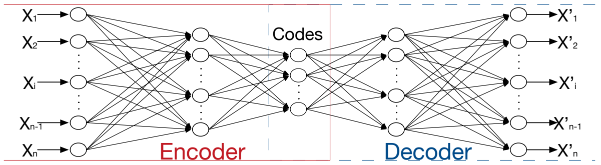

Figure 3. AE architecture.

2.2.2. Long Short-Term Memory

RNNs are often used to exploit the predictive capacity of temporal dependencies. However, these usually present problems of vanishing and exploiting gradients when dealing with very long sequences [8]. LSTM cells employ gates to avoid this problem and have become popular in recent years. Figure 4 shows all the components of an LSTM cell. Its mechanisms to remember relevant information are controlled by gates made up of ANNs with specific activation functions at the output layer. Thus, each one is responsible for filtering which information is relevant to the cell. This information is passed to the cell gate (horizontal line delimited by c_{t-1} and c_t in Figure 4). Two operations keep the relevant information. The forget gate, f_t, filters the information that the cell must forget. The second one is responsible for indicating what data are the new candidates to remember. In this way, the input gate, i_t, decides which values will be updated combined with new candidates, c'_t. This combination is added to the cell state. Finally, the output is a filtered version (tanh) of the cell state modulated.

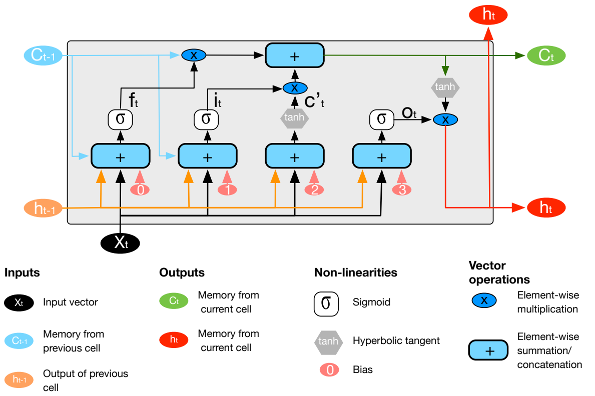

Figure 4. One cell Long Short-Term Memory RNN.

2.3. Metrics

The receiver operating characteristic (ROC) shows how the sensitivity and specificity of a binary classifier vary in terms of a detection threshold. The measure derived from this curve is the area under the ROC (AUROC), which takes values from 0 to 1, being 0.5 the case of a random classifier and 1 for the perfect one.

Another metric, used for the challenge is the utility score [1], which measures how well a model detect sepsis rewarding the early detection and penalizing the late/missed detections, this normalized metric takes values from 0 to 1, being 1 the perfect prediction and 0 the classifier with no positive predictions.

3. Results

Imputation: Laboratory tests are imputed taking a window of 6 hours of the variables. Then the 26 laboratory tests feed an AE with one hidden layer with 35 units and latent dimension (length of codes) of 15. For training the AE, ADAM optimizer with LR=0.001 is used. After parameter optimization, the minimum MSE was 0.039. This approach is compared with two classical imputation methods, i.e., imputation by the mean value and imputation by the last value.

Prediction: Vital signs, demographics, and information from the generated codes were merged to feed an LSTM, with input the evolution of 8 hours of each patient in ICU. Its target was the development of sepsis 6 hours before it occurred. The network had three hidden layers with 40, 30, and 25 units in each one, respectively. ADAM optimizer with a learning rate of 0.0001 and L2 regularization with $\beta = 0.0001$. For training the LSTM, 5-fold cross-validation was used. For training purposes as in [9],

Imputation method	AUROC	Utility
Mean	0.763	0.303
Forward filling	0.754	0.285
Proposed method	0.788	0.334

Table 1. Performance comparison common imputation methods vs proposed one in a subset of data (not official test scores).

Metric	Full test set	Test A	Test B	Test C
Utility	0.241	**0.344**	0.267	-0.247
AUROC	-	0.761	0.766	0.762

Table 2. Performance of methodology in hidden test sets (official test scores).

the extracted model from the fold that contains the best generalization for the patients was used.

In Table 1, the models using classic imputation have similar capacity. However, the proposed methodology has a better predictive capacity with a utility higher than 12% respect to the other imputation methods. Besides, in terms of dimensionality, using the codes, it was possible to reduce 27.5% the number of features to feed the LSTM model.

Finally, in Table 2, it can be appreciated the utility scores for the different hidden data sets. With these results, our team called *WIN-UAB* was ranked in the position 38/78 in the challenge, the utility ranges for all the teams were in the range [-0.841, 0.364]. Because the model does not know data from unit C, it does not generalize well in this data. However, for data whose part of its structure is known, generalization is adequate.

4. Conclusion

In this work, it was shown the potential of integrating and exploiting the predictive capacity of variables with few determinations. Thus, using the proposed imputation method together with DL techniques, it was possible to improve the early prediction of sepsis in ICU. On the other hand, it was possible to reduce dimensionality for the data using the codes of AE. The LSTM exploited the temporal dependencies of the stays of the patients in ICU. The results obtained in dataset C demonstrate that for unknown units, it is necessary to integrate a priori information so that the models can generalize in new ICUs. Finally, although the proposed mechanism does not present the best utility score, combining it with other machine learning approaches may have the potential to improve the early prediction of sepsis in the ICU and be truly useful in the clinical domain.

Acknowledgements

This research is supported by the Catalan Government under Project 2017 SGR 1670 and the Spanish Government under Project TEC2017- 84321-C4-4-R co-funded with European Union ERDF funds.

References

[1] 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.

[2] 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; ISSN 15354970.

[3] Macias E, Morell A, Serrano J, Vicario J. Knowledge extraction based on wavelets and dnn for classification of physiological signals: Arousals case. In 2018 Computing in Cardiology Conference (CinC), volume 45. IEEE, 2018; 1–4.

[4] 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;ISSN 15300293.

[5] Ford DW, Goodwin AJ, Simpson AN, Johnson E, Nadig N, Simpson KN. A Severe Sepsis Mortality Prediction Model and Score for Use with Administrative Data. Critical Care Medicine 2016;ISSN 15300293.

[6] Ibeas J, Macias E, Rubiella C, Morell A, Serrano J, Rodriguez-Jornet A, Vicario J, Rexachs D. Sp689 renal failure and mortality: From evidence to artificial intelligence, change of paradigm? Nephrology Dialysis Transplantation 2019;34(Supplement_1):gfz103–SP689.

[7] Singer M, Deutschman CS, Seymour C, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche JD, Coopersmith CM, Hotchkiss RS, Levy MM, Marshall JC, Martin GS, Opal SM, Rubenfeld GD, Poll TD, Vincent JL, Angus DC. The third international consensus definitions for sepsis and septic shock (sepsis-3), 2016.

[8] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation 1997;ISSN 08997667.

[9] Pisa I, Santín I, Vicario JL, Morell A, Vilanova R. ANN-based soft sensor to predict effluent violations in wastewater treatment plants. Sensors Switzerland 2019;ISSN 14248220.

Address for correspondence:

Edwar Macias
Telecommunications and Systems Engineering Department,
Univeritat Autònoma de Barcelona, 08193 Bellaterra, Spain
edwar.macias@uab.cat