Novel Imputing Method for the Early Prediction of Sepsis in ICU Using Deep Learning Techniques

E Macias, G Boquet, J Serrano, JL Vicario, J Ibeas, A Morell

Wireless Information Networking, Universitat Autònoma de Barcelona, Spain

Abstract

Using mechanisms that exploit the predictive capacity of data collected in intensive care units (ICU), it is possible to detect sepsis early. Vital signs are measured continuously, while laboratory test determinations are few. The combination of these sources of information generates a considerable number of missing values. In this manuscript is proposed some mechanisms to deal with data that have a high ratio of missing values, together with the use of deep learning techniques to improve the early detection of sepsis in ICU. Initially, the laboratory tests are separated and summarized. Then, the most representative information is extracted by taking codes from an autoencoder. This information is then combined with vital signals and used to exploit temporal dependence through long short-term memory recurrent neural networks. Finally, this method is compared to two classical imputation methods, i.e., imputation by the mean value and imputation by the last value. The proposed method has an outstanding performance with an area under the receiver operating curve of 0.79 and a utility (defined in the Early prediction of Sepsis from Clinical Data: the PhysioNet/Computing in Cardiology Challenge 2019) of 0.344, with a 30% dimensionality reduction compared to other imputation methods.

1. Introduction

In the era of big data and machine learning, it is possible to capture and extract knowledge from large amounts of clinical data. The evolution of patients is collected in electronic health records, where different registers such as images, physiological measurements or diagnoses, among others are stored. The follow-up of variables that describe the health condition of a patient depends on several factors such as the type of disease she/he suffers 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 phenomenon occurs in the intensive care unit (ICU), where vital signs are monitored continuously, while laboratory tests are taken less frequently. Combining several sources of information, missing values are generated for those determinations that do not match their timestamp. These are a type of the so-called missing not at random (MNAR) values [1]. On the other hand, one of the most critical problems in ICU is sepsis and its challenging early detection [2]. It represents an epidemiologic problem, with more than 30 million people who develop it and more than 6 million who die every year [3]. Although several works have started using machine learning [4–7] for the detection of pathologies, the most widely way to identify it is through clinical scores that relate the risk factors with events linearly [8]. However, the applications of other strategies to deal with the data 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 use both temporal evolution of patients and combining different sources of information, it is possible to improve the early detection of pathologies to support the clinical decisions. In this way, from massive data of patients and their progression in ICU, in this manuscript, it is proposed to combine a new mechanism to impute variables with high ratios of missing values with the application of deep learning (DL) techniques for the early detection of sepsis in the ICU. In summary, the main contributions of this manuscript are:

- Combine the information with different rates of missing values and impute in a novel way those variables that have high ratios of missing values.
- Extract the most relevant information from the data and reduce its dimensionality through the application of autoencoders.
- Exploit the predictive capacity of temporal evolution through the use of 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 ICU. 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. Figure 1 shows the methodology used in this work. Initially, due to the missing values problem, the sources of information are divided and processed separately. Then, they are combined and structured in order to feed an LSTM, which is in charge of exploiting the temporal dependencies in the data. 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.

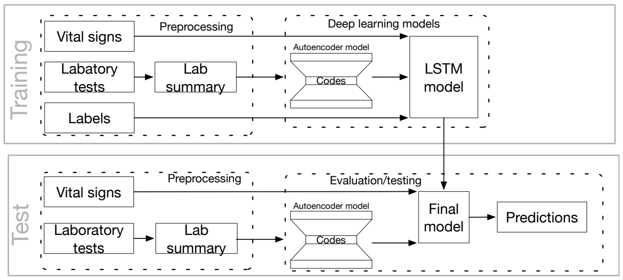

Figure 1. Work flow for the early prediction of sepsis in ICU

2.1. Preprocessing

Initially, the data is split randomly, into training and test sets (70-30%). Then, due to the different occurrence rates in laboratory tests and vital signs (see Figure 2), they are processed separately. Laboratory tests are summarized every $'N'$ hours to impute their missing values so that the tests during that time do not vary. Then, in order to extract the most relevant information from these data and decrease the dimensionality, codes of a trained autoencoder (AE) are extracted. In the case of vital signs, due they are monitored continuously, they are imputed using second-order interpolation.

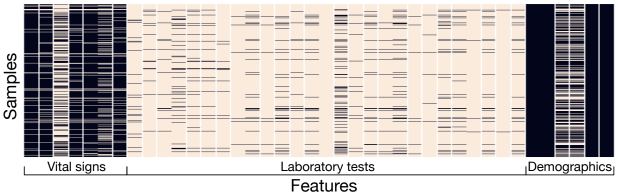

Figure 2. Missing values distribution for all the variables. Dark region refer to the available data

Finally, the two sets of variables are combined with the demographics, and these samples are structured to train a model based on LSTMs.

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(**W**)) for LSTM, Eq 1 and Eq 2 respectively.

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

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

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

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 through the minimization of the cost function iteratively. It is based on forward and back-propagation. In the first one, 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 to 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 accelerate 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 [9]. It uses first and second-order momentum to update the LR at each iteration. To avoid memorizing data in the training phase, a common problem on DL models, some techniques such as early stopping, increasing the dataset, applying regularizers, or eliminating network connections are applied. In this work, L2 regularization is used, which adds a term that penalizes the weights that tend to be very large and dropout to prevent the network from learning the training data through the random elimination of connections at each training interaction.

Thus, two DL models are used. One for representing the most significant information of laboratory tests in a smaller space, using the codes of anAE, and the second one to exploit the temporal evolution of patients in the ICU, through LSTM RNN.

2.2.1. Autoencoders

AEs 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 minimum error. 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 forces the AE to extract a complex structure, called codes, that best represents the data with a smaller dimension.

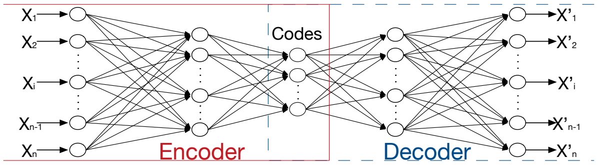

Figure 3. Autoencoder

In this work, the codes of a trained AE with the laboratory tests are extracted.

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 [10]. LSTMs 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 be able to remember relevant information are controlled by gates made up of ANNs with specific activation functions at the output layer. In this way, 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.

2.3. Metrics

The receiver operating characteristic (ROC) is used in this work to compare the different models. It 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.

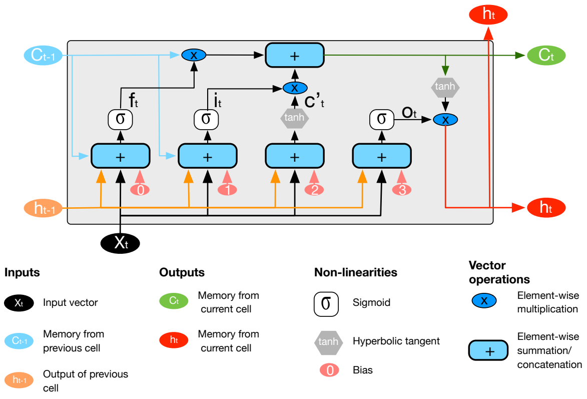
hₜ
Cₜ₋₁
Cₜ
fₜ
iₜ
C'ₜ
Oₜ
σ
tanh
x
+
0
1
2
3
hₜ₋₁
hₜ
Xₜ
Inputs
Xₜ Input vector
Cₜ₋₁ Memory from previous cell
hₜ₋₁ Output of previous cell
Outputs
Cₜ Memory from current cell
hₜ Memory from current cell
Non-linearities
σ Sigmoid
tanh Hyperbolic tangent
0 Bias
Vector operations
x Element-wise multiplication
+ Element-wise summation/ concatenation

Figure 4. Long Short-Term Memory RNN

Another metric, used for the *Physionet Challenge 2019-Early Prediction of Sepsis from Clinical Data* is the utility [2], 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.

3. Results

This methodology is compared with two conventional approaches used to handle missing values in medicine. First, by imputing the mean value of the variable, and the second one is imputing with the last determination of each variable. In both cases, variables not measured are imputed with the mean value of those variables for the training set. In the case of the proposed method, the laboratory tests are summarized every 12 hours. Then the 27 tests feed an AE with one hidden layer with 35 units and latent dimension (length of codes) of 15. For training, ADAM optimizer with LR=0.001, *tanh* as activation function and early stopping are used. After parameter optimization, the minimum MSE was 0.039.

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. The network had three hidden layers with 40, 30, and 25 units in each layer, respectively. *Tanh* as the activation function in their layers, ADAM optimizer with a learning rate of 0.00001, early stopping, L2 regularization with $\beta = 0.0001$ and dropout of 0.4 were used. For training the models, 5-fold cross-validation was used. For training purposes as in [11], the fold that contains the best generalization for the patients was used.

Imputation method	AUROC	Utility
Mean	0.763	0.303
Forward filling	0.754	0.285
Proposed method	0.788	0.344

Table 1. Performance comparison common imputation methods and proposed method

As it can be seen in Figure 5 and their respective utility in Table 1, the models using classic imputation have similar capacity. However, the proposed methodology has an outstanding predictive capacity with a utility higher than 12% respect to the other imputation methods. Besides, in terms of dimensionality, using the codes to represent the most significant information, it was possible to reduce 30% the amount of necessary data to feed the LSTM model.

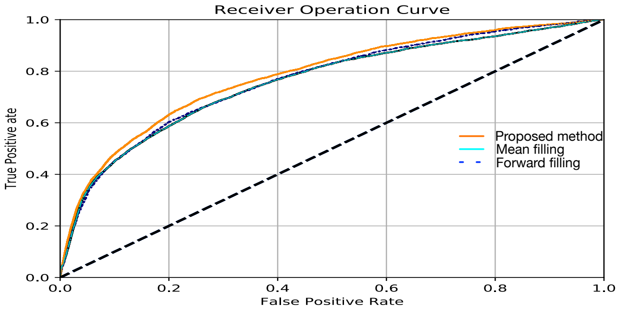

Figure 5. Performance comparison of three methods to deal with missing values

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 considerably. On the other hand, replicating information from the input to the output training an AE, complex structures were extracted in the codes in an unsupervised manner and was possible to reduce dimensionality for the data. Finally, the application of LTSM networks showed an outstanding predictive capacity with the three imputation methods, exploiting the temporal dependencies of the stays of the patients in ICU.

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] Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, Petersen I. Missing data and multiple imputation in clinical epidemiological research. Clinical Epidemiology 2017;ISSN 11791349.

[2] 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.

[3] Paoli CJ, Reynolds MA, Sinha M, Gitlin M, Crouser E. Epidemiology and costs of sepsis in the United States-an analysis based on timing of diagnosis and severity level. Critical Care Medicine 2018;ISSN 15300293.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv14126980 2014;.

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

[11] 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