POLITECNICO DI TORINO
Repository ISTITUZIONALE

Feature Selection Based on a Genetic Algorithm for Optimizing Weaning Success

Original
Feature Selection Based on a Genetic Algorithm for Optimizing Weaning Success / Rosati, S., Scotto, A., Fanelli, V., Balestra, G.. - 302:(2023), pp. 566-570. (Medical Informatics Europe 2023 Gothenburg (Sweden) 22-25 May 2023) [10.3233/SHTI230206].

Availability:
This version is available at: 11583/2979325 since: 2023-06-12T11:45:48Z

Publisher:
IOS Press - NLM (Medline)

Published
DOI:10.3233/SHTI230206

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

29 November 2026

Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)
© 2023 European Federation for Medical Informatics (EFMI) and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI230206

Feature Selection Based on a Genetic Algorithm for Optimizing Weaning Success

Samanta ROSATI[a], Andrea SCOTTO[a], Vito FANELLI[b] and Gabriella BALESTRA[a,1]
[a] *Department of Electronics and Telecommunications, Politecnico di Torino, Italy*
[b] *Department of Surgical Sciences, Università degli Studi di Torino, Italy*
ORCiD ID: Samanta ROSATI https://orcid.org/0000-0003-0620-594X
ORCiD ID: Andrea SCOTTO https://orcid.org/0000-0002-3317-3843
ORCiD ID: Vito FANELLI https://orcid.org/0000-0002-1647-2411
ORCiD ID: Gabriella BALESTRA https://orcid.org/0000-0003-2717-648X

Abstract. Finding the right time for weaning from ventilator is a difficult clinical decision. Several systems based on machine or deep learning are reported in literature. However, the results of these applications are not completely satisfactory and may be improved. An important aspect is represented by the features used as input of these systems. In this paper we present the results of the application of genetic algorithms to perform feature selection on a dataset containing 13688 patients under mechanical ventilation characterizing by 58 variables, extracted from the MIMIC III database. The results show that all features are important, but four of them are essential: 'Sedation_days', 'Mean_Airway_Pressure', 'PaO2', and 'Chloride'. This is only the initial step to obtain a tool to be added to the other clinical indices for minimize the risk of extubation failure.

Keywords. MIMIC, mechanical ventilation, weaning, genetic algorithms, feature selection

1. Introduction

Weaning from mechanical ventilation is the operation of liberating patients from respiratory support that ends with removing the endotracheal tube that is used to attach a patient to a ventilator in intensive care unit (ICU). Finding the right time for extubating a patient is a difficult clinical decision. Delayed weaning is associated with higher risk of patients' muscle atrophy, of nosocomial infections and with ICU's costs. Moreover, extubation failure increases the risk of patients' mortality.

Different support tools are available to estimate the probability of extubation failure (EF). The Rapid Shallow Breathing Index (RSBI) [1] is one of the most used indices for this purpose. However, it has high sensitivity but lower specificity. Different Machine Learning (ML) methods have been used by researchers to provide accurate EF estimation [2–5]. They have good results, but improvement is still needed.

The aim of our work is present some preliminary results related to the construction of a ML-based classifier to be used for the prediction of the weaning outcome. In particular, we focalize the attention on the selection of the most informative features for prediction purposes.

[1] Corresponding Author: Gabriella Balestra, E-mail: gabriella.balestra@polito.it.

2. Dataset Construction

2.1. Feature extraction

The dataset used in this study was extracted from the MIMIC III database [6]. First, from the available data we identified only patients that were supported by a ventilator during their stay in ICU. The number of identified patients was 23968. Then, four exclusion criteria were applied: age less than 18 years, death before extubation, duration of the ventilation lower than 24 hours, unintentional extubation. The result was an initial dataset containing 13691 patients.

For each patient we extracted a total of 91 variables describing personal (age, sex, weight, weight loss, height), and clinical information (SOFA score, GCS score, comorbidities, minimum, maximum and mean values of physiological parameters, laboratory parameters), weaning parameters, and the administered therapy. Patients were assigned to two classes: patients that were extubated and needed a new intubation within 48 hours (EF, Extubation Failure,'0') and patients who were successfully extubated and did not need reintubation within 48 hours (ES, Extubation Success,'1').

2.2. Data preprocessing

Data cleaning was performed to manage missing data and potential outliers. First, variables or patients with more than 70% of missing values were excluded as well as variables with many outliers. Then, for the remaining variables, we performed imputation of the missing values based on the kNN (k Nearest Neighbors) method, considering the patients of the two classes separately.

A further preprocessing step was applied to reduce the number of variables associated to the patient comorbidities. Starting from the initial 29 binary variables related to the comorbidities, we removed those variables equal to "0" for all patients included in our dataset. Then, we transformed the remaining binary variables into a decimal variable.

After data cleaning and preprocessing, we obtained a dataset containing 13688 patients and 58 variables. The dataset was highly imbalanced: the ES class contained 10832 patients whereas the other 2856 belonged to the EF class.

3. Feature Selection

3.1. Training set construction

The obtained dataset was divided into three sets. First, we randomly extracted 100 ES patients and 50 EF patients to be included in the Test Set (TS). For both groups, 50% of the patients were without imputed values.

We decided to use 2300 patients (1150 for each class) to construct a balanced Training Set (TRS). To this purpose, the remaining 13538 patients were divided according to the class and clustered using agglomerative hierarchical clustering. Then, we proportionally extracted, from the obtained clusters and for each class, 1000 patients containing imputed values and 150 without missing values. A Validation Set (VS) made of 11238 patients remained available to evaluate the feature selection results.

3.2. GA implementation

We chose the Deep Neural Network (DNN) as classification method to identify the weaning outcome. Dimensionality reduction was performed by means of feature selection to preserve the features meaning for explainability purposes. To optimize the subset of features to be used as input of the DNN, we decided to use a Genetics Algorithm (GA).

Each GA solution was made of 58 yes/no (binary) variables corresponding to the features, where 1 means the features is selected, 0 not selected.

The initial GA population consisted of 5000 solutions. At each iteration, 2000 parents were selected using the roulette wheel selection method. To each couple of parents was applied a crossover with 4 cutting points and probability equal to 1, followed by mutation with probability equal to 0.4.

The GA fitness function was based on the DNN classification results. After testing several DNN configurations with different numbers of hidden layers, we defined a triangle DNN structure in which the input layer consisted of a number of neurons equal to the number of selected features, the output layer was made of 1 neuron and the numbers of neurons of each hidden layers was progressively decreased by 4 with respect the previous layer.

The fitness function, to be minimized, was computed using the following formula:

$$fitness = 1 - \frac{TP+TN}{TP+TN+FP+FN} + 0.3 * \left| \frac{TP}{TP+FN} - \frac{TN}{TN+FP} \right| \quad (1)$$

where TP, TN, FP and FN are respectively true positives, true negatives, false positives and false negatives. This fitness equation aimed to prefer solutions with high accuracy but, at the same time, to avoid too much imbalance between sensitivity and specificity.

For each solution, we trained a DNN using the TRS and selecting only the subset of features identified in the current solution. Then, we used the trained network to classify the VS elements and calculate the corresponding fitness value. If the percentage of patients correctly classified was lower than 10%, the current network was discarded and the training was repeated up to a maximum of 5 times. If all the trained DNNs returned unsatisfactory results, the fitness value was set to 1000 to be sure to discard the actual solution in the next generations.

4. Results

4.1. Data cleaning results

The 91 variables were checked to find missing values. From this analysis it emerged that 35 variables had no missing values, 34 variables had less than 1% of missing values, 13 variables had a number of missing between 1% and 40%, 5 variables had missing values between 40% and 70%, and only 4 variables had more than 70% of missing values. The number of missing was decreased removing 3 patients that presented a huge number of variables with no value associated. Moreover, we decided to remove the 4 variables with more than 70%. We also removed height and weight with 32% and 17% of missing values respectively, because the imputation of these variables could be affected by a high

uncertainty and could introduce biases in the classifier construction. After data cleaning, the number of remaining variables was 85. The number of outliers was negligible.

4.2. GA results

Due to the stochastic nature of the algorithm, we decided to perform 50 repetitions of GA starting from the same initial population. For each repetition, we saved the solution with the best (lowest) fitness value. Analyzing the entire set of 50 best solutions, the number of select features ranged from 21 to 39, with a median value of 30. Figure 1 shows, for each variable, how many times it was selected. Four variables were included in most of the solutions (≥40 times): 'Sedation_days', 'Mean_Airway_Pressure', 'PaO2', and 'Chloride'. All variables were selected at least 8 times in the 50 solutions, meaning that all of them are somehow informative in combination with others for the weaning outcome prediction.

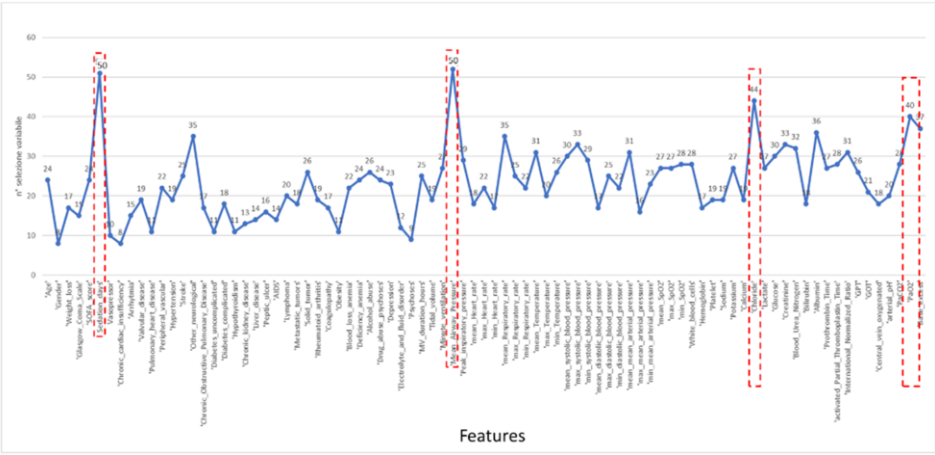


Figure 1. This diagram reports the number of solutions containing each variable.

To assess the goodness of the 50 GA solutions, we analyzed the classification results obtained using the corresponding DNN. We did not perform any tuning on the DNN implementation. Figure 2 shows sensitivity and specificity obtained, on the training and validation sets, using each best solution. Specificity is, on the average, slightly higher than sensitivity considering the TRS (0.72±0.05 vs 0.68±0.06, respectively), whereas on the VS the results are similar (0.69±0.05 vs 0.68±0.06 for specificity and sensitivity respectively). This behavior is essentially due to the penalty term inserted in the fitness function that allows to prefer solutions with classification performance balanced between the two classes. Moreover, it can be observed that the performances on the TRS and VS are quite similar, highlighting the generalization capability of the classifier.

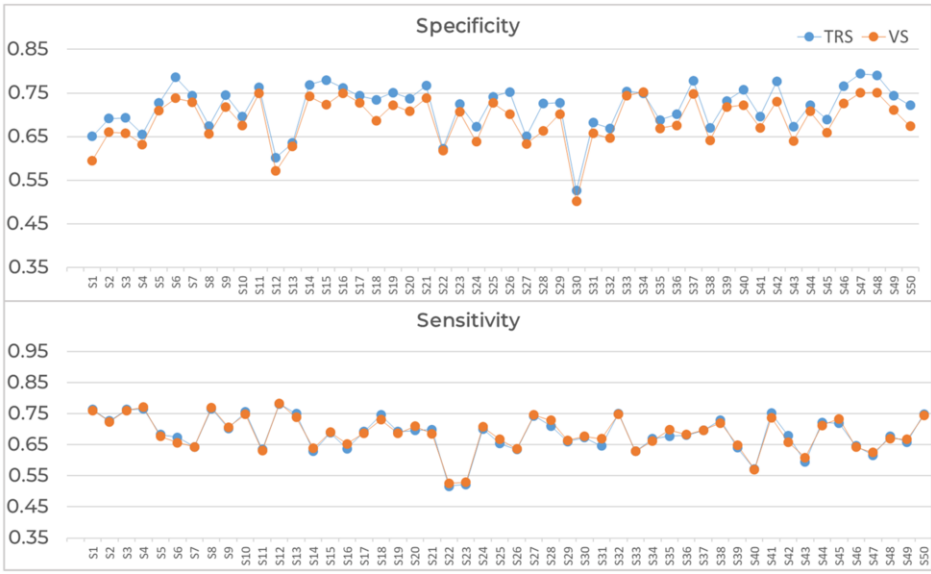


Figure 2. Sensitivity and specificity associated to each GA best solution.

5. Conclusions

In this work we focalized on the role of different kinds of features to predict the outcome of weaning from a mechanical ventilator. Looking at how many times each variable occurs in the best GA solutions, we can state that all features are somehow informative, but four of them are almost essential for this purpose. It is also important to notice that there are different groups of variables associated with them giving similar results. This is only an initial step to obtain a tool to be added to the other clinical indices for minimize the risk of extubation failure.

References

[1] Karthika M, Al Enezi FA, Pillai L V, Arabi YM. Rapid shallow breathing index. Ann Thorac Med 2016;11:167–76. https://doi.org/10.4103/1817-1737.176876.

[2] Fabregat A, Magret M, Ferré JA, Vernet A, Guasch N, Rodríguez A, et al. A Machine Learning decision-making tool for extubation in Intensive Care Unit patients. Comput Methods Programs Biomed 2021;200:105869. https://doi.org/https://doi.org/10.1016/j.cmpb.2020.105869.

[3] Getzen E, Ungar L, Mowery D, Jiang X, Long Q. Mining for Equitable Health: Assessing the Impact of Missing Data in Electronic Health Records. J Biomed Inform 2023:104269. https://doi.org/https://doi.org/10.1016/j.jbi.2022.104269.

[4] Tsai T-L, Huang M-H, Lee C-Y, Lai W-W. Data Science for Extubation Prediction and Value of Information in Surgical Intensive Care Unit. J Clin Med 2019;8. https://doi.org/10.3390/jcm8101709.

[5] Zeng Z, Tang X, Liu Y, He Z, Gong X. Interpretable recurrent neural network models for dynamic prediction of the extubation failure risk in patients with invasive mechanical ventilation in the intensive care unit. BioData Min 2022;15:21. https://doi.org/10.1186/s13040-022-00309-7.

[6] Johnson A, Pollard T, Mark R. MIMIC-III Clinical Database (version 1.4). PhysioNet. 2016. Available from: https://doi.org/10.13026/C2XW26.