
Ines Krissaane¹, Richard Wilkinson ¹, Kingsley Hampton ², Jumanah Alshenaifi³

¹ School of Mathematics and Statistics, University of Sheffield, Sheffield, UK
² IICD Department, University of Sheffield, Sheffield, UK
³ MD Andreson UTHealth Graduate School of Biomedical Sciences, Houston, TX, USA

Abstract

Sepsis is one of the leading causes of morbidity and mortality in hospitals. In this paper we propose a machine learning approach for anomaly detection to aid the early detection of sepsis. Using the medical data of over 40,000 patients [1], we use both unsupervised and supervised methods to extract relevant features from the data, and then use standard classification approaches to predict sepsis six hours before clinical diagnosis occurs. To extract features, we use the reconstruction error of an auto-encoding neural network trained on health patients, and use random forest classifiers to learn the most important features for the classification of patients. We then combined the features from both of these approaches with a variety of standard classification models. Cross-validation is used to evaluate the resulting models. We find the best classification utility function (of 0.245) is achieved with a logistic regression classifier.

1. Introduction

Machine learning techniques have been used previously to better understand sepsis [2] and to improve sepsis prediction [3]. As a contribution to the Physionet/Computing in Cardiology Challenge 2019 [1], this paper focuses on the implementation of a modelling pipeline using ML techniques to detect patients with sepsis earlier that it is actually declared in ICU.

1.1. Sepsis condition

Sepsis is defined as a life-threatening organ dysfunction caused by a dysregulated body’s response to infection [2]. Sepsis follows a continuum through severe sepsis to septic shock, starting with the systemic inflammatory response syndrome (SIRS). The early detection of sepsis is highly valuable and may potentially save a patient’s life [5], reduce medical complications, and reduce the cost to the healthcare system [6]. A patient is diagnosed SIRS if they meet at least two of the following criteria:

- Temperature > 38 °C or < 36 °C
- Heart rate > 90/min
- Respiratory rate > 20/min or PaCO2 < 32 mm Hg
- White blood cell count > 12000/mm³ or < 4000/mm³.

By interrogating a large dataset of hospitalized patients, our aim is to identify which clinical measurements can best be used to predict sepsis, and moreover, to train models to predict the probability a given patient will develop sepsis.

1.2. Anomaly detection

Sepsis detection can be seen as an anomaly detection problem where you may think about an individual who developed sepsis as a patient with specific abnormal representative features, i.e., for each patient, we aim to detect a change from their usual clinical measurements. We use data provided as part of the Computing in Cardiology Challenge 2019 [1] which contains clinical data from 40,336 ICU patients, of which 2,932 develop sepsis. The aim is to distinguish between the usual variation in patient data that occurs in ill patients who don’t have sepsis, with the change in measurements that occur in patients who develop sepsis. Our approach is to use a selection of standard classification methods, trained using carefully learned features based on two machine learning methods, random forests (RF) and auto-encoding neural networks (AENN), both of which have proven to be effective at extracting important features, and at identifying irregularities in data.

1.2.1. Supervised learning problem

A random forest [7] consists of multiple random decision trees, where each individual tree gives a class prediction. To classify a new instance, the forest chooses the classification having the most votes over all the trees in the forest. As a classifier, random forest performs an implicit feature selection, using a small subset of variables for the classification. And Gini importance can be used to indicate how often a particular feature was selected for a split [8]. We will therefore consider the whole dataset and will extract the most important features.
1.2.2. Unsupervised learning problem

Sepsis is detected in less than 2% of the patients in our data. Since abnormality is a low probability event, data from the abnormal class are rare and detection can be targeted as an unsupervised learning problem. Unsupervised anomaly detection techniques detect anomalies in an unlabelled data set under the assumption that the majority of the instances in the data set are healthy patients, as in this case. Autoencoders [9] are a type of artificial neural network which learn to copy its inputs to its outputs, via a hidden layer of limited size. The representation of an input in its hidden layer can then be seen as an efficient compression of the data. We can use them in an anomaly detection scenario [10] considering a neural network trained only on the normal patients. For a new instance, we expect the reconstruction error to be higher for abnormal patients.

To formulate the problem, we introduce some notation. The \( N \) patients in the EHR database are modeled as a set of pairs \( D = \{(x_1, y_1), \ldots, (x_N, y_N)\} \), where the (hidden) label \( y \in \{1, +1\}^T \) is a vector of sepsis condition labels (+1 denoting sepsis) indicating at each measurement point whether a patient has sepsis or not. The clinical measurements \( X \) is a matrix of 40 different clinical measurements over time \( x_{t,j} \) where \( 1 \leq t \leq T \) and \( 1 \leq j \leq 40 \).

2. Materials and Methods

2.1. Data pre-processing and correlation analysis

For each patient, we used 40 clinical measurements over time, including demographics, vital signs, and laboratory values from where we eliminated any variable which had greater than 80% missing values, leaving just 13 variables. In addition, the variables Unit1 and Unit2 are equally distributed across the whole dataset, and are uncorrelated with the sepsis label \( y \), and were therefore removed. \( SBP \) and \( DBP \) are highly correlated with \( MAP \) via the relationship \( MAP = (SBP + 2*DBP) / 3 \), and so were also discarded as well as \( HospAdmTime \). For the 8 remaining variables, we impute any missing values by filling gap with the non missing values forward or backward along a Series. We merged all our patient files and add an extra column \( ID \) for the patient identifier. The variable \( O2Sat \) corresponds to the pulse oximetry percentage, and would be controlled for ventilated patients. Our methods find that \( O2Sat \) is not appearing as a predictive variable (See Fig. 3), and we did not include this variable in the analysis. The remaining seven variables are given in the Table 1, and the data processing pipeline is described in Fig. 1.

### Table 1. List of the seven variables included in the analysis.

<table>
<thead>
<tr>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Respiration Rate</td>
</tr>
<tr>
<td>Mean Arterial Pressure</td>
</tr>
<tr>
<td>Gender</td>
</tr>
<tr>
<td>Heart Rate</td>
</tr>
<tr>
<td>ICULOS (Hours spent in ICU)</td>
</tr>
</tbody>
</table>

2.2. Random forest importance features computation

We want to evaluate the importance of the remaining features in the prediction of sepsis. We split the dataset into training and test sets in proportion 70:30. Using the training data, we perform a 3-fold cross-validation grid search to tune the hyper parameters of the RF. The optimal values for the RF were as follows: the number of trees \( n_{estimators} = 200 \) in the forest, the maximum number of features \( max\_features = 7 \) considered for splitting a node, the minimum number of data points \( min\_samples\_split = 2 \) placed in a node before the node is split and the minimum number of data points \( min\_samples\_leaf = 1 \) allowed in a leaf node. The Gini importance was used to calculate feature importance from the RF [8].

2.3. Auto-encoding neural networks

We train a dense feed-forward neural network auto-encoder using only the patients who do not have sepsis. During training, we minimize the \( L_2 \) reconstruction error (RE), which is the mean squared distance between input and output: \( L(X, X') = \|X - X'\|^2 \). We use the seven selected features from the RF as inputs, and the neural network architecture used, composed of only dense layers (7/56/392) is described in Fig. 2. We use the auto-encoder by passing patient data through it and evaluating the reconstruction error for all points. This will be an additional feature for the classification.

2.4. Classification and Cross-Validation

Six standard classifiers were used in this study (see Table 2) with and without the RE’s feature. Five-fold cross validation was used to evaluate each classifier, performance evaluated with the expected utility (the Computing in Cardiology Challenge defined an asymmetric utility function which heavily penalizes false negative predictions). Logistic regression was found to be the classifier...
with the highest expected utility (\(=0.245\)).

Data processing, feature extraction, and classification were all performed using Python 3.6. We used Keras in Tensorflow for the AENN and the library Scikit-Learn for classification. All the implementation is available in: https://github.com/ineskris/SepsisChallenge-Cinc2019

3. Results

3.1. Feature importance

Fig. 3 shows the relative importance of each feature as estimated by the random forests and the Gini importance. As might be expected, the variable Age is the most significant in line with previous work [11][12]. Sepsis is equally prevalent in both genders, and so as expected the Gender variable is judged to be of low importance. The variable O2Sat will not be use in the AENN.

3.2. Reconstruction error

Fig. 4 shows the observed distribution of reconstruction errors evaluated using a test set of 1,000 patients, half of who have sepsis. Observe that the RE for those with sepsis is higher on average than for those without sepsis. Precision and recall are commonly used to evaluate the accuracy of anomaly detection problem.

In Fig. 5, the precision/recall shows how the trade off between missing an abnormal patient and the cost of falsely flagging a patient who is healthy according to different threshold for the RE. Taking a \(RE = 4000\) makes totally sense with Fig. 4.

3.3. Classification Model

Table 2 highlights the positive impact of the RE on the prediction accuracy for all classifiers. Using the expected utility function from the challenge and not the accuracy, the logistic regression classifier gave predictions with the highest score.

4. Conclusion

Anomaly detection methods enabled us to identify from a patient’s data unusual patterns which do not conform to
expected behaviour. In this paper, we used features extracted from the RF and a new feature, the RE obtained from the AENN, to predict sepsis with a classical logistic regression model. The method was validated on PhysioNet Challenge 2019 dataset and the results are encouraging, suggesting that the early prediction of sepsis is an achievable task (Utility Score = 0.245). We intend to further develop our method namely the auto-encoder part to find a more appropriate architecture for the neural network.

References


Address for correspondence:
Ines Krissaane
School of Mathematics and Statistics
The University of Sheffield, UK
ikrissaane1@sheffield.ac.uk