Early Prediction of Sepsis Using Random Forest Classification for Imbalanced Clinical Data

Simon Lyra[1], Steffen Leonhardt[1], Christoph Hoog Antink[1]

[1] Medical Information Technology, Helmholtz-Institute for Biomedical Engineering, RWTH Aachen University, Germany

Abstract

The early prediction of sepsis in intensive care units using clinical data is the objective of the PhysioNet/Computing in Cardiology Challenge 2019. *In this paper, a machine learning approach is presented which uses an optimized Random Forest for prediction of a septic condition. After an initial data augmentation step, a customized learning process is performed for the trees to consider imbalance in the dataset. Finally, a feature reduction is implemented and the forest is trimmed to 50 trees for an optimal classification in terms of run time and accuracy. Using a 10-fold cross-validation on the complete training dataset, a mean utility score of 0.376 is achieved. In the final submission, a normalized observed utility score of 0.296 on the full test set is achieved. Our team name is The Septic Think Tank (final rank: 21).*

1. Introduction

Every year about 6 million people die from the consequences of resulting dysfunctions and multiple organ failures due to a sepsis [1]. Although sophisticated surveillance systems which use electronic clinical data (e.g. MEWS, qSOFA) are applied, a correct and accurate prediction of a disease onset remains elusive [2]. Several studies have demonstrated that machine learning tools can decrease diagnostic uncertainties and identify septic patients by accessing clinical data in real-time. In [3] a dynamic Bayesian network was implemented to model the progression of organ failure. Furthermore, neural networks have been used for state classification [4] and a sepsis progression estimator was implemented using support vector machines [5].

To encourage the development of new algorithms for an early prediction of sepsis, the PhysioNet/CinC Challenge 2019 aims to address the problem by providing clinical data from intensive care units [6]. In this work, a machine learning algorithm based on the Random Forest (RF) classifier is described.

2. Methods

For classification, an ensemble of decision trees as implemented in the MATLAB `TreeBagger` function is used. This approach evaluates the results of many individual trees, which reduces overfitting and improve the generalization [7]. During both learning and prediction phase, not-a-number (NaN) values are treated as individual missing information, i.e. a data point or a variable is only omitted completely if *all* values are NaN. In the development phase, 10-fold cross validation was used for algorithm optimization. For submission, the complete available dataset was used for training and the resulting algorithm was submitted for evaluation on the hidden test sets.

2.1. Features and Feature Reduction

All 40 features from the dataset were used for prediction without further preprocessing. In addition, at each time step t, mean and standard deviation of the previous t_Δ time steps were calculated. Thus, a total of $N_{\mathrm{feat}} = 120$ candidate features were calculated. To reduce them, a straightforward elimination process was applied: In the prediction step, each feature was once replaced by NaN and the "normalized observed utility" (NOU) was calculated. Thus, a feature's importance is expressed by its reduction of NOU.

2.2. Data Augmentation

Without data augmentation, the dataset is severely imbalanced: pooling $N_{\mathrm{rec}} = 40,330$ recordings of the training data (six recordings were discarded to allow for 10-fold cross validation with an equal number of $N'_{\mathrm{rec}} = 4,033$ patients), $N_{\mathrm{neg}} = 1,524,071$ timesteps belong to the non-septic state, while $N_{\mathrm{pos}} = 27,916$ belong to the septic state. According to the rules of the challenge, a positive utility is achieved if sepsis is detected up to 12 hours before its actual onset and at max 3 hours late. Thus, all timesteps within this range are labeled positive. After this procedure, a slightly less imbalanced dataset is obtained with $N'_{\mathrm{neg}} = 1,509,992$ and $N'_{\mathrm{pos}} = 41,995$ respectively.

 ISSN: 2325-887X DOI: 10.22489/CinC.2019.276

2.3. Imbalanced Random Forest Learning

After data augmentation, a ratio of $r = N'_{\text{neg}}/N'_{\text{pos}} = 35.96$ is obtained. To allow for unbiased learning, the following strategy is employed: In each iteration, $N_0 = 10$ trees are learned using all N'_{pos} positive data points as well as an equal number of randomly selected negative data points. The process is repeated $\lfloor r \rfloor = 35$ times until (almost) all data points are used. All trees are aggregated to form a forest with $N_{\text{trees}} = \lfloor r \rfloor N_0 = 350$ trees.

2.4. The *Sashiki* Forest

To enforce causality, evaluation is performed using a for-loop in the official phase of the challenge. The MATLAB `TreeBagger` implementation shows a large overhead when not executed in "bulk-mode": Predicting a randomly generated dataset with 40 features and 100 timesteps, we obtained a run-time of 3.8 seconds if the `predict` function is called with some `TreeBagger` object and a 100 by 40 feature matrix. If `predict` is called with the same `TreeBagger` object but using a timestep-by-timestep for-loop, i.e. executed 100 times with a single feature vector, the whole prediction process took 288.8 seconds. Thus, the need to drastically reduce computational time arose. Moreover, reducing the ensemble's complexity might facilitate the understanding of its decision process.

To optimize the RF, the following algorithm was developed. First, each tree is used individually to predict the complete dataset. This leads to a prediction matrix $\mathbf{P}_{i,j}$ with the dimension $N_{\text{timesteps}} \times N_{\text{trees}}$ containing a "1" if tree j decides sepsis for timestep i and "0" otherwise. Next, the sum of each row is calculated to obtain the unnormalized base prediction vector

$$\hat{P}_i^0 = \sum_{j=1}^{N_{\text{trees}}} \mathbf{P}_{i,j}. \tag{1}$$

Now, for each tree j, a prediction vector P^j is calculated, in which that specific tree is left out of the prediction process. This vector is normalized by the number of remaining trees,

$$P_i^j = \frac{\hat{P}_i^0 - \mathbf{P}_{i,j}}{N_{\text{trees}} - 1}. \tag{2}$$

Note that using this approach, only one subtraction and one division is necessary for each tree and timestep once the base prediction vector is calculated, significantly reducing computational time. Next, the quality of prediction $q(j)$ is evaluated for each left-out tree using a pre-defined threshold p_{th}, the ground-truth Y and an evaluation function $F(\cdot)$,

$$q(j) = F\left(Y, P^j > p_{\text{th}}\right). \tag{3}$$

Finally, the tree j_{opt} is identified whose omission led to the maximum quality value,

$$j_{\text{opt}} = \arg\max_j q(j) \tag{4}$$

and omitted in the next iteration. The process is repeated with $N_{\text{trees}} \rightarrow N_{\text{trees}} - 1$ and an updated $\mathbf{P}$ until only one tree is left.

The obvious choice for $F(\cdot)$ is the supplied evaluation function that calculates the NOU from [6]. However, it is fairly computational expensive and has to be executed $N_{\text{trees}} \cdot N_{\text{trees}}/2$ times in the tree-elimination process, rendering it impractical for algorithm development. In the 2015 challenge [8], a modified accuracy function

$$\text{MA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + w \cdot \text{FN}} \tag{5}$$

was used for scoring, where TP indicate true positive and TN indicate true negative predictions. The parameter $w = 5$ was used to particularly penalize false negative (FN) over false positive (FP) predictions of arrhythmia alarms. In the current challenge, the same general concept can be applied, since an undetected sepsis (i.e. a FN prediction) also severely influences the NOU (and of course patient outcome in a real application). In Figure 1, an overview of the submitted algorithm is illustrated.

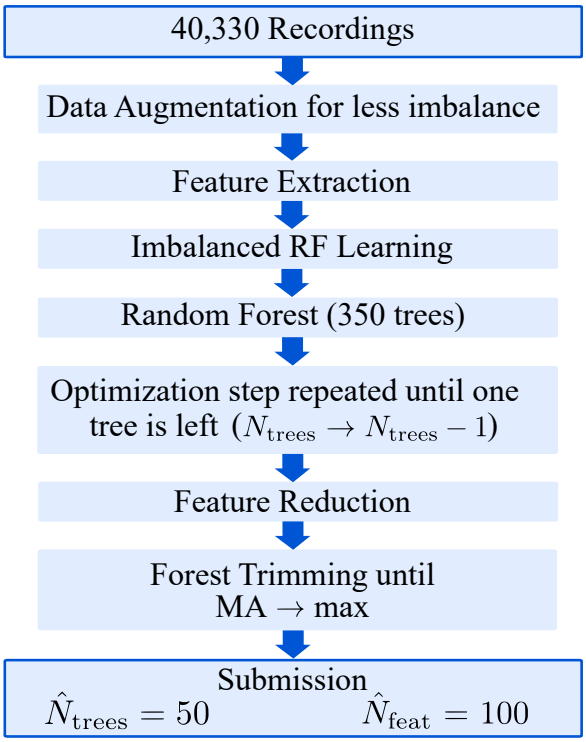

Figure 1. Overview of the submitted algorithm.

3. Results and Discussion

3.1. Features and Feature Reduction

Using cross validation, $t_\Delta = 4\,\mathrm{h}$ was found to achieve optimal results. Figure 2 shows the features ranked by importance in terms of NOU reduction. As can be seen, the replacement of the 20 most important features (lowest rank) leads to a sharp decrease in NOU, while the replacement of the 20 least important features (highest rank) seems to increase it overproportionally. Thus, a subset of $\hat{N}_{\mathrm{feat}} = 100$ features (gray area) was used in the final implementation.

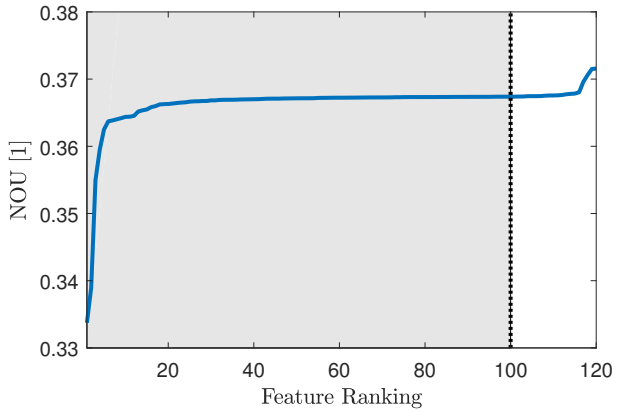

Figure 2. Features ranked by importance. A lower rank indicates that the replacement of a feature with NaN values decreases NOU more than a higher-ranked feature.

3.2. Trimming the Forest

First, the penalization parameter w was manually optimized to obtain a computational-efficient surrogate for the NOU. For this, the base prediction vector is normalized

$$P^0 = \frac{\hat{P}^0}{N_{\mathrm{trees}}}. \tag{6}$$

Next, the threshold p_{th} is varied from 0.4 to 0.7 and both, NOU and MA with variable w are calculated for the predictions $P^0 > p_{\mathrm{th}}$. To optimize w, correlation of NOU and MA is maximized. Figure 3 shows NOU and MA for $w \in 34, 44, 54$. Standardization is applied for better visual comparison. As can be seen, NOU and MA match best visually for $w = 44$. Additionally, a correlation coefficient of $\rho = 0.9955$ is obtained while the calculation of MA is about 5800 times faster compared to the MATLAB-implementation of NOU.

Thus, tree reduction is conducted with MA and $w = 44$. Figure 4 shows the results of the tree reduction procedure. As expected, MA initially increases as more and more trees

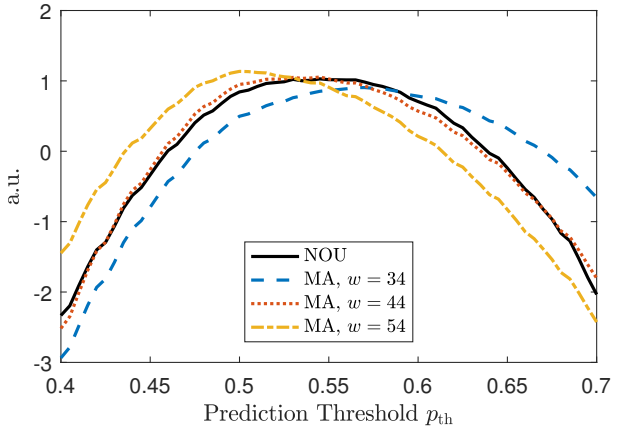

Figure 3. Comparison of Normalized Observed Utility (NOU) and Modified Accuracy (MA) for different values of the penalty parameter w over the prediction threshold p_{th}. Curves are standardized for better visual comparison.

are eliminated, i.e. as the number of trees is reduced. MA plateaus for a number of approximately 50 to 100 trees. If the number of trees is reduced beyond 25, a sharp drop in MA is observed. Thus, in the final implementation, a total of $\hat{N}_{\mathrm{trees}} = 50$ trees is used.

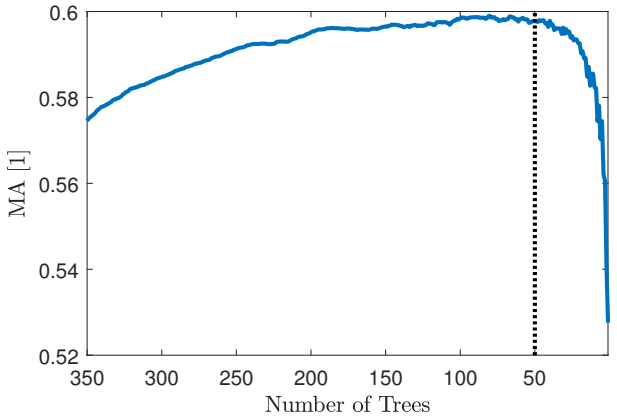

Figure 4. MA of the tree reduction process with $w = 44$.

In Table 1 the results of a 10-fold cross-validation on the complete (augmented) dataset are shown. For the evaluation, a mean NOU of 0.376 with a standard deviation of 0.023 is achieved. For the full hidden test set, a NOU of 0.296 was achieved for the final entry of our team *The Septic Think Tank* (final rank: 21). One can see that the results for the test sets A and B do not deviate strongly from the 10-fold cross-validation. This indicates a generalization process during the learning phase, so an overfitting of the RF was prevented. Additionally, the results provide the evidence that the hidden test sets A and B do not vary significantly from the training data set. In contrast to this, a large

deviation can be observed for test set C. Here, significantly varying features are assumed, which strongly influence the prediction result of the RF.

Table 1. Results for 10-fold cross-validation and submission results on the complete dataset.

Fold	AUROC	AUPRC	Accuracy	f	NOU
1	0.787	0.086	0.832	0.117	0.353
2	0.812	0.080	0.847	0.117	0.344
3	0.819	0.092	0.851	0.132	0.401
4	0.813	0.104	0.846	0.126	0.385
5	0.829	0.091	0.852	0.131	0.413
6	0.807	0.103	0.846	0.119	0.365
7	0.821	0.113	0.877	0.145	0.400
8	0.795	0.097	0.874	0.128	0.348
9	0.804	0.087	0.832	0.119	0.367
10	0.816	0.097	0.844	0.127	0.388
Mean	0.810	0.095	0.850	0.126	0.376
SD	0.012	0.010	0.014	0.008	0.023
Set A	0.788	0.083	0.808	0.121	0.372
Set B	0.828	0.089	0.892	0.132	0.378
Set C	0.780	0.043	0.730	0.041	-0.218
Full					0.296

Finally, it can be seen that a RF classifier is an appropriate approach for the prediction of a septic condition from clinical data. However, due to incomplete, imbalanced data and the often nonspecific and subtle symptoms in the disease progression, it is very hard to achieve a robust and satisfying result for an early prediction.

4. Conclusion and Outlook

In this paper an approach for early prediction of sepsis using RF classification was presented. In general, it has been observed that the algorithm is capable of predicting a septic condition. However, due to the low specificity and large variety of septic symptoms, it is very hard to find the most significant features for a robust prediction.

Several measures to improve the classification result can be taken. During data augmentation, no interpolation algorithm was used to fill the missing values in the data, so the decision trees are trained with incomplete information. In a former version of the algorithm, a forward insertion was performed where missing values were set using the previous data row, but no considerable improvement was observed, so the interpolation was neglected. Nonetheless, there are more sophisticated approaches to predict missing data points which could improve the result.

In the future, the RF approach could be enhanced by using a more complex selection of combined features, e.g. the ratios of pulse rate, respiration, blood pressure, etc.. Feature combinations have the potential to gain supplementary information from the interdependencies of the clinical data. Additionally, a more detailed analysis of feature importance could be performed. By setting a focus on the parameters described in MEWS, camera-based unobtrusive measurement techniques as introduced in [9] are sufficient for a patient-friendly surveillance system. Finally, a contact-free vital sign monitoring could decrease disease onsets which are caused by the sensors itself (nosocomial infections).

5. Acknowledgements

The authors gratefully acknowledge financial support provided by the German Research Foundation [Deutsche Forschungsgemeinschaft, LE 817/26-1, LE 817/32-1].

References

[1] Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, Suppes R, Feinstein D, Zanotti S, Taiberg L, Gurka D, Kumar A, Cheang M. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Critical Care Medicine jun 2006;34(6):1589–1596.

[2] W van der Woude S, F van Doormaal F, A Hutten B, J Nellen F, Holleman F. Classifying sepsis patients in the emergency department using sirs, qsofa or mews. The Netherlands journal of medicine 05 2018;76:158–166.

[3] Peelen L, de Keizer NF, de Jonge E, Bosman RJ, Abu-Hanna A, Peek N. Using hierarchical dynamic bayesian networks to investigate dynamics of organ failure in patients in the intensive care unit. Journal of Biomedical Informatics apr 2010; 43(2):273–286.

[4] Brause R, Hamker F, Paetz J. Septic Shock Diagnosis by Neural Networks and Rule Based Systems. Heidelberg: Physica Verlag HD. ISBN 978-3-7908-1788-1, 2002; 323–356.

[5] Wang SL, Wu F, Wang BH. Prediction of severe sepsis using SVM model. In Advances in Experimental Medicine and Biology. Springer New York, 2010; 75–81.

[6] Reyna M, Josef C, Jeter R, Shashikumar S, M. Brandon Westover M, Nemati S, Clifford G, Sharma A. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. Critical Care Medicine 2019; (in press).

[7] Breiman L. Random forests. Machine Learning Oct 2001; 45(1):5–32.

[8] Clifford GD, Silva I, Moody B, Li Q, Kella D, Shahin A, Kooistra T, Perry D, Mark RG. The PhysioNet/computing in cardiology challenge 2015: Reducing false arrhythmia alarms in the ICU. In 2015 Computing in Cardiology Conference (CinC). sep 2015; .

[9] Brueser C, Hoog Antink C, Wartzek T, Walter M, Leonhardt S. Ambient and unobtrusive cardiorespiratory monitoring techniques. IEEE Reviews in Biomedical Engineering 2015; 8:30–43.

Address for correspondence:
Simon Lyra (lyra@hia.rwth-aachen.de)
Medical Information Technology
Helmholtz Institute, RWTH Aachen University
Pauwelsstr. 20 / D-52074 Aachen / Germany