Uncertainty-Aware Model for Reliable Prediction of Sepsis in the ICU

Marco AF Pimentel[1], Adam Mahdi[1], Oliver Redfern[2], Mauro D Santos[1], Lionel Tarassenko[1]

[1]Department of Engineering Science, University of Oxford, Oxford, UK
[2]Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK

Abstract

Predicting the onset of sepsis from routinely collected clinical data is challenging, as physiological and laboratory measurements are sampled at different frequencies and missing data are not randomly distributed. We propose a two-model approach, where the first model predicts a probability of sepsis and the second estimates the uncertainty of these predictions. We then optimize a "decision rule", which considers both the probability and uncertainty to make the final prediction.

A range of features was derived from the original time patient records at each time point, including demographics, the most recent recorded (non-null) value for each vital sign and laboratory result, maximum and minimum values of each vital sign within the preceding 12 and 24 hours, and difference between the two last recorded values of a subset of laboratory results. This feature set was used to train a Gradient Boosting Machine (GBM) classification model to predict sepsis (within 6 hours). A second GBM regression model was used to estimate the uncertainty of those predictions using a different set of features based on the time elapsed since the last recorded value of each vital sign and laboratory tests at each time point. Optimal hyperparameters for both models were determined using Bayesian optimisation with a 5-fold cross validation (70% records from each training set). The outputs from both models were then combined using logistic regression (using 15% of records available) to calculate a re-calibrated probability of sepsis. The combined model was evaluated on the held-out test set (15% of records available) using the area under the receiver-operating characteristics curve (AUROC) and the Utility score.

Our uncertainty-aware approach achieved an AUROC of 0.830 and a Utility score of 0.421 on the held-out test set. The Utility score of our model is substantially higher than the baseline model supplied with the current challenge (Utility score of 0.180) when evaluated on the held-out test set.

We have developed a novel prediction approach which considers uncertain estimates when predicting the onset of sepsis. The proposed model performs favourably on the held-out data set.

1. Introduction

Sepsis is defined as "life-threatening organ dysfunction caused by dysregulated host response to infection". The early detection and treatment of sepsis can lead to better patient outcomes [1, 2]. With the significant growth in the uptake of Electronic Health Record (EHR) systems, there has been a surge in the number of studies developing prediction models to aid the identification of sepsis in ICU patients. EHR data, however, pose a challenge to standard approaches to using machine learning to model longitudinal data.

Most data sets derived from routinely collected clinical data contain a substantial proportion of missing values and irregularly-sampled data. In particular, some laboratory tests are only performed in a subset of patients for diagnostic purposes (e.g. troponin). Furthermore, even vital signs and common laboratory tests (e.g. renal function) are measured at different intervals (which may range from 1 hour to 72 hours) and vary across patients. Most approaches for coping with these missing values or irregularly-sampled data rely on imputation methods, without accounting for the potential biases (and errors) that it may generate when making these predictions.

We propose a two-model approach, where the first model predicts a probability of sepsis and the second estimates the uncertainty of these predictions due to missing data. We then optimise a *decision rule*, which considers both the probability and model's uncertainty to make our final predictions.

2. Materials and methods

This study was performed as part of the Physionet Challenge 2019 [3].

2.1. Dataset

This study used two datasets provided for the Physionet Challenge 2019. These datasets were originally extracted

by the Challenge's coordinators from patient admissions to three (sets A, B, C) intensive care units (ICU). All datasets contained hourly-stamped measurements for 34 distinct (time-varying) variables (8 vital signs and 26 laboratory test results), including the time (or hour) at which each value was gathered. In additions, the values of 5 demographic (static) variables are also available for each record in the datasets. Two datasets (set A and set B) with a total of 40,336 ICU patient records were made available for model development, with corresponding sepsis labels (1 if onset of sepsis occurs within 6 hours for patients who developed sepsis, 0 otherwise) provided for each hourly-stamped measurement in each record. The third dataset (set C) was not available to the Challenge participants but was used to evaluate the final models during the Computing and Cardiology 2019 conference.

2.2. Data pre-processing

Prior to feature extraction, the distribution of each physiological measurement was manually inspected. Physiologically implausible values for certain variables are set as missing (null entries). Specifically, this process was performed for heart rate, respiratory rate, systolic and diastolic blood pressures, mean arterial pressure, temperature and pulse oximetry.

2.3. Statistical analysis

Our method relies on a two-model approach: (1) the first model estimates the probability of sepsis using an augmented set of features derived from the clinical data available; (2) the second model attempts to estimate the uncertainty (or error) of the predictions of the first model generated by the imputation method. For the latter, we use a second set of features that relate to the "missingness" of each time-varying variable. Both prediction and model's uncertainty are then combined to provide a re-calibrated probability of sepsis.

For the Physionet Challenge 2019, the official metric used to assess the performance of the submitted models is a customized Utility score, which rewards early prediction of sepsis (up to 12 hours before onset) and penalizes late predictions [3].

2.4. Feature extraction

Each patient's risk of sepsis is computed for every time point, and our model considers the whole sequence of physiological and laboratory measurements recorded up to that timestamp. Thus, we converted each record's hourly-stamped set of variables into a new set of hourly-stamped variables based on the previous and current measurements available for that record.

For the first model, a range of extracted and derived features was extracted from the original time series. Static variables (e.g. age, gender), were simply repeated at each time point. For time-varying variables, this was done by extracting the most recent measured value for each vital sign and laboratory measurement, maximum and minimum values of each vital sign within the preceding 12 and 24 hours, and difference between the two last recorded values of a subset of laboratory results, including creatinine, blood urea nitrogen (BUN), and platelet count (these were set to 0 if no two previous values were available for a given variable). If a variable was completely missing for a given patient (or there are no previous or current values), the median value over the training data for that variable was imputed.

For the second model, we computed the elapsed time (in hours) since the last recorded (non-null) value for each of the 34 time-varying variables. These features capture whether a given variable was measured at a given timestamp and provides information about how much time has passed since the last measurement. If a variable is completely missing for a given record, or there are no previous or current values, a value of 1 year (8760 hours) was imputed.

2.5. Model description

Both models are based on Gradient Boosting Machines (GBMs). The GBM is an ensemble method based on using weak learners, in our case, decision trees. GBM iteratively trains collections of decisions trees to classify the training data; with each step incorporating a new decision tree, which preferentially weights the correct classification of previously misclassified training examples. We chose a GBM method on the basis of favourable comparison with other regression-based methods we have considered, and due to the XGBoost implementation [4], which provides options for regularization and the handling of imbalanced classes.

XGBoost provides many hyperparameters to control both the entire ensemble and individual decision trees. To prevent overfit of the models to the training data, we used early stopping, which allows to stop the training (i.e., adding more trees) when validation scores have not improved for 50 iterations.

2.6. Hyperparameter tuning

Given the vast combination of hyperparameters to explore and their domain (i.e., the range of values that we want to evaluate for each hyperparameter), we used Bayesian optimization. Bayesian hyperparameter optimization finds the value that minimizes an objective function by building a surrogate function (probability model) based on past evaluation results of the objective. The surrogate is cheaper to optimize than the objective, so

the next input values to evaluate are selected by applying a criterion to the surrogate (in our case, the expected improvement was used). Bayesian methods differ from random or grid search in that they use past evaluation results to choose the next values to evaluate.

Hyperparameters of both models were tuned using 5-fold cross validation of the training set using the area under the receiver operating characteristics curve (AUROC) for scoring the first model, and the root-mean-squared-error for the second model. A tree Parzen estimator was used as the optimization algorithm.

The domain of each hyperparameter tuned (same for both models) is shown in Table 1.

Table 1. GBM's hyperparameters domain and distribution. Names of hyperparameters are those used in the XGBoost package (see [4]).
* *n_estimators* is fixed as it is estimated via early stopping.

Hyperparameter	Domain	Sampling
*n_estimators**	10,000	Fixed
eta	[0.001, 1.0]	Log-Uniform
max_depth	[2, 9]	Uniform
subsample	[0.4, 1.0]	Uniform
colsample_bytree	[0.1, 0.8]	Uniform
gamma	[0.1, 5.0]	Uniform
scale_pos_weight	[0.2, 20.0]	Uniform
lambda	[0.1, 3.0]	Uniform

All other hyperparameters were set to their default value.

2.7. Model development and assessment

We evaluated the performance of the proposed prediction model by randomly dividing the 40,336 patient records into a training set (containing 70% of records from set A, and 70% of records from set B), a recalibration set (containing 15% of records from each set), and a testing set (containing the remaining 15% of records from each set).

First, we used the training set to train a GBM (binary) classification model to predict sepsis within 6 hours of a given timestamp (the "Sepsis Label") using the feature set and the best set of hyperparameters found using the hyperparameter tuning procedure described in the previous section. Secondly, we computed the negative log-likelihood for each prediction in our training set, and trained a second GBM regression model to estimate the error of those predictions using the second feature set and the best set of hyperparameters found for this second model.

Using the recalibration set, we then determined the predicted values from both models (i.e., the probability of sepsis, and the model's uncertainty) for each record's entry, and combined both predictors into a single recalibrated score (our *decision rule*) with logistic regression, using "sepsis label" as the outcome. Finally, in order to provide a binary prediction of "sepsis", it was necessary to threshold the score value (a probability between 0 and 1). The threshold was that which maximized the Utility score on the recalibration set.

We report the performance results of the proposed approach for the held-out testing set (which was not used at any point during the development of the model presented in this study). We compare the performance of our proposed model with that of the baseline model supplied with the current Challenge. We also compare the performance of our uncertainty-aware approach with that our first model (the threshold was re-calculated for this model using the same methodology).

All data processing, model development and assessment, and statistical analyses were performed using Python.

3. Results

The evaluation metrics on the held-out testing set (containing 6051 records from set A and set B) for each model are shown in Table 2. Evaluation metrics include the AUROC, the area under the precision-recall curve (AUPRC), the F1-score, the accuracy and the Utility score.

Table 2. Evaluation metrics of the uncertainty-aware model (our two-model approach, M2), the single GBM model that does not include the estimations of model uncertainty (single-model approach, M1), and the baseline model based on a time-to-event regression model (B0).

Metric	B0	M1	M2
AUROC	0.710	0.826	**0.830**
AUPRC	0.053	0.094	**0.099**
Accuracy	0.771	0.800	**0.826**
F1-score	0.076	0.113	**0.122**
Utility score	0.180	0.391	**0.421**

Our uncertainty-aware approach achieved an AUC of 0.830 and a Utility score of 0.421. The Utility score of our model is substantially higher than the baseline model (Utility score of 0.180) when evaluated on the held-out testing set.

4. Discussion and conclusions

At the thresholds (or operating points) found for each model, we observed a large improvement in the Utility score with our uncertainty-aware model (M2) relative to the baseline model, B0 (0.421 vs. 0.180), and a smaller, yet substantial, improvement relative to the single model

(M1), which does not include the uncertainty estimates (0.421 vs. 0.391).

An important goal that we aimed to achieve with modelling the uncertainty of the GBM prediction model was achieving higher reliability in prediction. Prediction reliability is orthogonal to prediction accuracy, and a study [5] showed that state-of-the-art machine learning models are not reliable as they are not well-calibrated to correlate model confidence with model strength. Thus, we evaluated our uncertainty calibrated model (M2) against the GBM model with no uncertainty recalibration (M1), and the results Table 2 show that, although the improvement in discrimination (as given by the AUC) is relatively modest (0.830 vs. 0.826), the improvement of the Utility score is substantial (0.421 vs. 0.391). Hence, the predictions from the uncertainty-aware model appears to provide better calibrated predictions.

We proposed an uncertainty-aware approach that has the potential to enhance reliability of both interpretations and predictions of sepsis provided by a GBM model. Further analysis of prediction reliability may be necessary in order to demonstrate that the model is accurately calibrated and thus can defer predictions when making prediction with "I don't know" option.

References

[1] M. Singer et al., "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)", JAMA, vol. 315, no. 8, pp. 801–810, Feb. 2016.

[2] C. W. Seymour et al., "Time to Treatment and Mortality during Mandated Emergency Care for Sepsis", *N Engl J Med*, vol. 376, no. 23, pp. 2235–2244, Jun. 2017.

[3] M. Reyna et al., "Early prediction of Sepsis from Clinical Data – the PhysioNet Computing in Cardiology Challenge 2019", *https://physionet.org/content/challenge-2019/1.0.0/*, Aug. 2019.

[4] T. Chen, and C. Guestrin, "XGBoost: A scalable tree boosting system", *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016.

[5] M. .P. Naeini, G. F. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using Bayesian binning", *AAAI*, Jan. 2015.

Address for correspondence:

Marco AF Pimentel
Institute of Biomedical Engineering
Department of Engineering Science
University of Oxford
Old Road Campus Research Building, Roosevelt Dr
Oxford OX3 7DQ, UK
E-mail: *marco.pimentel@eng.ox.ac.uk*