The Signature-based Model for Early Detection of Sepsis from Electronic Health Records in the Intensive Care Unit

James Morrill[1,3], Andrey Kormilitzin[1,2], Alejo Nevado-Holgado[2], Sumanth Swaminathan[3], Sam Howison[1], Terry Lyons[1]

[1] Mathematical Institute, University of Oxford, Oxford, United Kingdom
[2] Department of Psychiatry, University of Oxford, Oxford, United Kingdom
[3] Iterex Therapeutics, 500 7th Ave, New York, NY 10018, United States

Abstract

Optimal feature selection leads to enhanced efficiency and accuracy when developing both supervised and unsupervised machine-learning models. In this work, a new signature-based regression model is proposed to automatically identify a patient's risk of sepsis based on physiological data streams and to make a positive or negative prediction of sepsis for every time interval since admission to the intensive care unit. The gradient boosting machine algorithm that uses the features at the current time-points and the signature features extracted from the time-series to model the longitudinal effects of sepsis yields the utility function score of 0.433. This is the highest scoring entry from 417 submissions in the official phase of the challenge. The signature method shows a systematic and competitive approach to model sepsis by learning from health data streams.

1. Introduction

Sepsis is thought to be present in more than half of hospitalisations that lead to death in the US [1]. Early detection of sepsis would likely have profound consequences on hospital mortality rates. For example, it is widely reported that mortality rates increase significantly for each hour of delay in receipt of antibiotics [2].

Here we propose a signature-based, machine learning approach to generating a risk score that a given patient will develop sepsis using hourly averaged patient data from the time since admission to the current time. The data and performance metrics used are those set out in the PhysioNet Challenge 2019 [3]. The method takes in data sequentially and uses the signature transformation to turn the time-series data into useful features. These features are fed, along with the variable information at the current time point, into a gradient boosting algorithm to learn combinations of features relevant to sepsis, which then leads to a risk score for the patient.

2. Methods

2.1. The Signature of a Path

A path X of finite length in d dimensions can be described by the mapping $X : [a, b] \to \mathbb{R}^d$, or in terms of co-ordinates $X = (X_t^1, X_t^2, ..., X_t^d)$, where each coordinate X_t^i is real-valued and parametrised by $t \in [a, b]$. The signature transformation S of a path X is defined as an infinite collection of terms:

$$S(X)_{a,b} = (1, S(X)^1_{a,b}, S(X)^2_{a,b}, ..., S(X)^d_{a,b}, S(X)^{1,1}_{a,b}, S(X)^{1,2}_{a,b}, ...), \quad (1)$$

where each term is a k-fold iterated integral of X with multi-index $i_1, ..., i_k$:

$$S(X)^{i_1,...,i_k}_{a,b} = \int_{a<t_k<b} ... \int_{a<t_1<t_2} \mathrm{dX^{i_1}_{t_1}} ... \mathrm{dX^{i_k}_{t_k}}. \quad (2)$$

Similarly to statistical moments of a d-dimensional vector-valued random variable, such as mean, variance and higher moments, one can define statistical moments of a *path*-valued random variable, which are essentially the signature moments [4] defined in (2). Path-valued random variables are naturally observed in many problems involving ordered events, such as time-series data, patient-generated measurements, speech etc. The signature $S(X)$ completely characterises a path X up to tree-like equivalence and is invariant to reparameterisation [5]. The usefulness of signatures as features of sequential data was demonstrated theoretically for non-parametric hypothesis testing [4] and algebraic geometry [6] as well as in numerous machine learning applications [7], for example: in healthcare [8–12], finance [13, 14], computer vision [15, 16], topological data analysis [17] and deep signature learning [18].

2.2. Implementation

2.2.1. Sepsis Labels

The challenge data is labelled with the value '1' for patients who develop sepsis where $t \geq t_{sepsis} - 6$ and 0 for $t < t_{sepsis} - 6$, t_{sepsis} being the time of sepsis onset as defined by the Sepsis-3 definition. For patients who never develop sepsis the data is labelled zero everywhere. Predictions are scored for their binary classification performance against a utility function described fully in the challenge description. False positives are penalised in non-septic patients and zero score is given for true negative predictions. For septic patients, early prediction is penalised, false negative predictions are more heavily penalised, and true positive predictions yield a positive score.

Given that we are optimising the utility score, not simply the percentage of correct binary predictions, we create a labelling that takes into account information about the utility score. We define the utility value, U, as the difference in utility score from predicting a 1 over a 0. That is, suppose the prediction of a 0 gives score U_0 and predicting 1 gives score U_1, then we label the sample with $U = U_1 - U_0$. This labelling gives larger absolute values (they can also be large and negative) for the samples that lead to a larger absolute utility score and are thus more important to label correctly. We build a regression model that takes this labelling as the target variable.

2.2.2. Data Imputation

Missing values during a patient's hospital stay are imputed by using a forward-fill method. If no previous value exists the value is left as 'NaN'.

2.2.3. Hand-Crafted Features

The signature features are augmented by a number of extra features from the 40 physiological measures from the dataset. These features and their definitions are listed in Table 1. Two new features 'ShockIndex', which is defined as the heart rate divided by the systolic blood pressure and 'BUN/CR' which is the ratio of levels of bilirubin to creatinine, were introduced following the work [19] where the authors showed their importance in sepsis detection. As sepsis is labelled only if there has been a 2-point deterioration in 'SOFA' score within a 24-hour period, an auxiliary 'PartialSOFA' score is additionally introduced. The partial 'SOFA' score is constructed from the variables that present both in the 'SOFA' score and our dataset. Additionally, the indicator variable SOFA_deterioration is marked as '1' if 'PartialSOFA' saw this deterioration over the last 24 hours.

The laboratory and temperature features are sparsely filled due to measurements being infrequently taken. We hypothesise that measurement frequency will give an indication of condition progression and severity, since measurements are likely to be taken more frequently when doctors are more concerned with patient health. We thus include a feature that counts the number of measurements that have been taken over some given look-back window.

Finally, we include the maximum and the minimum value of each vital sign variable over some look-back window.

Table 1. Features derived from the data.

Feature name	Description
ShockIndex	Heart Rate / Systolic Blood Pressure
BUN/CR	Bilirubin / Creatinine
PartialSOFA	Score of the SOFA components that are found in the challenge data
SOFA_Deterioration	Binary label given 1 if PartialSOFA has increased by 2 in the last 24 hours

2.2.4. Signature Features

To extract longitudinal information from the timeseries, we turn to the signature transformation as outlined in Section 2.1. A sliding window approach is used so that the signature features are computed for each time-point over a window of some given look-back size. The signatures of 'PartialSOFA', 'MAP' and 'BUN/CR' are computed with a time dimension and the lead-lag transformation and then signatures of all non-stationary columns are computed after first applying the cumulative sum followed by the lead-lag transformation. For information on the individual transformations see [8]. The signature truncation levels and look back window sizes are treated as hyperparameters and found during model optimisation.

2.3. Hyperparameters

The following hyperparameters were used in model training (here signature refers to both the basic signature features and the signature features after a cumulative sum transform has been applied, as each takes the same hyperparameters, though they can have different values): The look-back windows for the count variable, the signatures and the min/max computation, the features we compute the signature of, the truncation level of the signature (the signature order), and the classifier hyperparameters. Table 2 lists each of these with the value used in training the final model.

2.3.1. Model training and validation

We use the stratified 5-fold cross validation method where the folds are chosen to contain approximately the

Table 2. List of all hyperparameters with the values used for each in the final model.

Parameter	Final Value
Count look-back	8
Sig look-back	7
CSig look-back	7
Min/max look-back	6
Sig columns	Partial_SOFA, BUN/CR, MAP
CSig columns	All non-stationary
Sig order	3
CSig order	3
Sig leadlag	True
CSig leadlag	True

Sig = Signature, CSig = Cumulative sum signature

same number of time points and septic cases. No patient has data in more than one fold. We use the light gbm implementation of gradient boosting regression [20] as our algorithm to regress against our modified sepsis labels. We then use a gradient-free optimisation algorithm to determine the cutoff threshold on the regressed values that maximises the utility score on the training set. The regressor and threshold are applied to the validation set to make our overall predictions and evaluate the score.

3. Results

The cross validation scores on each of the 5-folds of the training data are given in Table 3. The model achieved a 0.433 score on the hidden test set, higher than our averaged score on the training set meaning the model is likely quite robust and has not been subject to much overfit.

Table 3. Scores on each cross validation fold.

	Fold					
	1	2	3	4	5	**Average(std)**
Utility score	0.432	0.434	0.437	0.448	0.4	0.43(0.018)

3.1. The Usefulness of Signatures

Table 4 displays the average 5-fold CV scores on the training set from models trained on different subsets of features giving us an idea on the predictive power of the additional features. Inclusion of the signature values gives good improvement on the overall utility score: the signature transformation is successfully uncovering relevant information from the time-series that can be used to discriminate cases of sepsis.

Table 4. Scores from models trained on different sets of features

Features	Averaged utility score
Time only	0.282
Original 40 features only	0.389
Hand-crafted features included	0.418
Hand-crafted features and signatures included	0.430

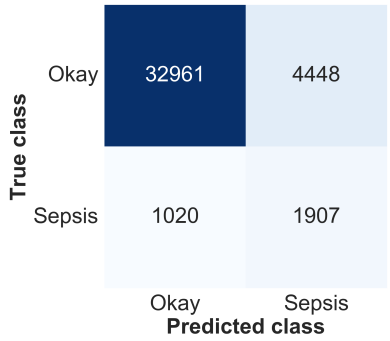

Figure 1. Confusion matrix displaying the number of people predicted as likely to get sepsis compared with those who actually end up with sepsis with the threshold tuned to 30% specificity.

3.2. Early Detection of Sepsis

Whilst the goal of the challenge was to optimise the score achieved on the pre-defined utility function, it is useful to consider how the model can be used in an in-hospital environment to provide clinically actionable information. Given the output of the regressor, we select a threshold such that once exceeded, the patient is marked 'at risk' of developing sepsis. This threshold can be chosen to achieve clinically meaningful sensitivity and specificity. The AUC ROC value when considering a varying threshold for this early detection compared against the people who actually develop sepsis is 0.868. As an example, setting the threshold such that a 30% sensitivity is achieved results gives a confusion matrix as displayed in Figure 1. At this specificity, 65.3% of sepsis cases are identified correctly. Of these 67.5% are predicted early (before the desired 6 hour window), 14.2% are predicted in the desired window and 18.3% are late. This shows that whilst the model can be used as an effective screening tool for sepsis, it does not in general predict cases in the desired 6-hour window before they occur, it generally predicts much further in advance.

4. Discussion

We have presented a signature-based model for early prediction of sepsis. We showed that the signature representation produced a useful summary of the longitudinal physiological measurements that was used to effectively

discriminate septic from non-septic cases. The addition of the signature terms improved significantly the predictive algorithm as demonstrated in Table 4. The method proposed has achieved the highest score on the utility function in the official phase of the challenge where there have been over 400 submissions. We have also shown that the model predictions can be turned into clinically actionable information for use by doctors. We saw that patients could be labelled as sepsis risk patients with an AUC of 0.868 score considering the score against those who do eventually develop sepsis.

Acknowledgements

AK is supported by the MRC under the Pathfinder programme grant MC/PC/17215. JM is supported by the EPSRC under the program grant EP/L015803/1 in collaboration with Iterex Therapeutics. TL is supported by the EPSRC under the program grant EP/S026347/1 and by the Alan Turing Institute under the EPSRC grant EP/N510129/1. TL, JM and AK are members of the DATASIG programme.

We are thankful to Scott Ganis for stimulating discussions throughout this challenge.

References

[1] Rhee C, Jones TM, Hamad Y, Pande A, Varon J, OBrien C, Anderson DJ, Warren DK, Dantes RB, Epstein L, Klompas M, for the Centers for Disease Control and Prevention (CDC) Prevention Epicenters Program. Prevalence, Underlying Causes, and Preventability of Sepsis-Associated Mortality in US Acute Care Hospitals. JAMA Network Open February 2019;2(2):e187571. ISSN 2574-3805.

[2] 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 June 2006;34(6):1589–1596. ISSN 0090-3493.

[3] 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.

[4] Chevyrev I, Oberhauser H. Signature moments to characterize laws of stochastic processes. arXiv preprint arXiv181010971 2018;.

[5] Hambly B, Lyons T. Uniqueness for the signature of a path of bounded variation and the reduced path group. arXiv preprint math0507536 2005;.

[6] Pfeffer M, Seigal A, Sturmfels B. Learning paths from signature tensors. SIAM Journal on Matrix Analysis and Applications 2019;40(2):394–416.

[7] Chevyrev I, Kormilitzin A. A primer on the signature method in machine learning. arXiv preprint arXiv160303788 2016;.

[8] Kormilitzin A, Saunders K, Harrison P, Geddes J, Lyons T. Application of the signature method to pattern recognition in the cequel clinical trial. arXiv preprint arXiv160602074 2016;.

[9] Kormilitzin A, Saunders KE, Harrison PJ, Geddes JR, Lyons T. Detecting early signs of depressive and manic episodes in patients with bipolar disorder using the signature-based model. arXiv preprint arXiv170801206 2017;.

[10] Gligic L, Kormilitzin A, Goldberg P, Nevado-Holgado A. Named entity recognition in electronic health records using transfer learning bootstrapped neural networks. arXiv preprint arXiv190101592 2019;.

[11] Arribas IP, Saunders K, Goodwin G, Lyons T. A signature-based machine learning model for bipolar disorder and borderline personality disorder. arXiv preprint arXiv170707124 2017;.

[12] Moore P, Lyons T, Gallacher J, Initiative ADN, et al. Random forest prediction of alzheimers disease using pairwise selection from time series data. PloS one 2019; 14(2):e0211558.

[13] Arribas IP. Derivatives pricing using signature payoffs. arXiv preprint arXiv180909466 2018;.

[14] Kalsi J, Lyons T, Arribas IP. Optimal execution with rough path signatures. arXiv preprint arXiv190500728 2019;.

[15] Yang W, Lyons T, Ni H, Schmid C, Jin L, Chang J. Leveraging the path signature for skeleton-based human action recognition. arXiv preprint arXiv170703993 2017;.

[16] Xie Z, Sun Z, Jin L, Ni H, Lyons T. Learning spatial-semantic context with fully convolutional recurrent network for online handwritten chinese text recognition. IEEE transactions on pattern analysis and machine intelligence 2017; 40(8):1903–1917.

[17] Chevyrev I, Nanda V, Oberhauser H. Persistence paths and signature features in topological data analysis. IEEE transactions on pattern analysis and machine intelligence 2018; .

[18] Bonnier P, Kidger P, Perez Arribas I, Salvi C, Lyons T. Deep signatures. arXiv preprint arXiv190508494 2019;.

[19] Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Science Translational Medicine August 2015; 7(299):299ra122–299ra122. ISSN 1946-6234, 1946-6242.

[20] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. LightGBM: A Highly Efficient Gradient Boosting Decision Tree ;9.

Address for correspondence:

Andrey Kormilitzin
Mathematical Institute, Woodstock Road, Oxford, OX2 6GG
andrey.kormilitzin@maths.ox.ac.uk