A Comparison of Machine Learning Tools for Early Prediction of Sepsis from ICU Data

Po-Ya Hsu¹, Chester Holtz¹

¹ University of California, San Diego, La Jolla, USA

Abstract

We explore the efficacy of modern machine learning methods for the task of modeling sepsis progression. We applied a novel imputation and feature selection scheme based on signal processing technology and our medical expertise. We compared the performance of several approaches including neural networks, sparse quantile regression, and baseline classification algorithms such as random forest and SVMs. We conclude that the application of neural network, random forest, sparse quantile regression, neighborhood algorithms, and naive Bayes classifiers yields superior performance with respect to accuracy, sensitivity, and specificity.

1. Introduction

Sepsis is a life-threatening condition which occurs when a host’s response to infection results in severe damage to its own organs and tissue [1]. The classification, identification, and treatment of sepsis is well studied, however, the disease remains a serious health issue with high rates of morbidity, and even in-hospital mortality [2, 3]. Early detection of Sepsis is vital for effective treatment, and each hour left untreated increases the chance of death, especially in the ICU [4, 5]. The task of detecting sepsis early is typically formulated as a multi-channel temporal classification task. Clinical data is commonly sampled irregularly, thus often requiring a set of hand-crafted preprocessing steps, such as binning, carry-forward imputation, and rolling means prior to the application of a predictive model. However, such naive imputation schemes lead to a loss of data sparsity, which may carry crucial information in this context. In light of these challenges, we propose to leverage spline-based interpolation models to imputation of unobserved samples and normalize the timescale of the variables.

1.1. Prior Work

Effective models for early sepsis prediction are desired since sepsis is a life-threatening condition. According to [4, 5], each one-hour delay of treatment of a case leads to an increase of \( \sim 7\% \) in the mortality rate. The 2019 PhysioNet Computing in Cardiology (CinC) challenge seeks to develop automated methods for early sepsis detection based on the ICU data.

There is a maturing body of prior work on data-driven sepsis detection methods applying machine learning algorithms [6-8]. In [7], random forest method is shown to generate promising results in conjunction with vital signs and lab values. In [6] and [8], unique machine learning approaches have been developed to achieve outstanding sepsis prediction. [9] evaluate two techniques for early prediction of sepsis: a temporal convolutional network, and a KNN-based approach leveraging Global Alignment Kernels. Our work is also similar to that of [10] who adopt a neural network classifier augmented with a multi-task Gaussian process regression layer to interpolate vital signs.

1.2. Contribution

We evaluated several baseline machine learning classification algorithms and deep learning techniques in this study. To resolve the issue of missing and imbalanced data, we developed an imputation and feature selection scheme for ICU data. In summary, our contributions include:

- We proposed a method to handle irregularly sampled data via a spline-based imputation algorithm.
- We evaluated the performance of a variety of statistical machine learning algorithms on engineered features.
- We proposed a novel deep learning-based video classification framework for the task of sepsis prediction.

2. Method

In this section we review the technical details of the algorithms we applied to this problem and detail our analysis and numerical results. We decompose the prediction problem into stages as in Fig [1]. We evaluated several different classes of predictors on our selected and learned features. In this section, we will briefly summarize and discuss the relative performance of each algorithm we applied and address their advantages and disadvantages in the context of
sepsis detection.

2.1. Formulation

We frame the problem of early detection of sepsis as a multivariate time series classification problem. Given a new patient encounter, the goal is to continuously update the predicted probability that the encounter will result in sepsis, using all available information up until that time. We follow the general framework from [10]. Given a dataset $D = \{d_i\}_{i=1}^N$ consisting of $N$ independent patient encounters, each patient encounter $d_i$, is described by a set of covariates and covariate vectors. Baseline covariates $b_i \in \mathbb{R}^{B \times 1}$ are available on hospital admission and consist of demographic information including gender and age. Lab covariates $l_i \in \mathbb{R}^{L \times t}$ and vital-sign covariates $v_i \in \mathbb{R}^{V \times t}$ are observed online for each time-step $0 \leq t \leq T$ where $t = 0$ denotes hospital admittance time. The goal is to predict, for each $t$ the likelihood that a particular patient $d_i$ has or will contract sepsis.

It is important to note that $T$ may vary over patients in the dataset, and that $l_i$ and $v_i$ are only partially observed due to the irregular sampling procedure employed for each clinical variable. Additionally, each encounter in the training set is associated with a binary vector $o_i \in \{0, 1\}^T$ denoting whether or not the patient has acquired sepsis at each time step. Thus, the data for a single patient encounter can be summarized as a 4-tuple: $d_i = \{b_i, l_i, v_i, o_i\}$. For brevity, we adopt the notation $d'_i = \{b_i, l'_i, l_i, o_i\}$ to mean the observed sequence of variables at time $t$.

2.2. Data preprocessing and experimental setup

The training dataset for the Challenge consisted of 40,336 subjects. For each subject, the data included demographics, vital signs, laboratory values, onset time of sepsis, and sepsis label. We evaluated several approaches to preprocess the data. In particular, the prevalence of missing values and a tailed sequence-length distribution were the primary issues that affected the design of our preprocessing pipeline.

We train our method with k-fold cross validation. 80% of the full dataset, setting aside 10% as a validation set to select hyperparameters and a final 10% for testing.

2.3. Imputation and normalization

Prior to classification with our machine learning-based models, we logarithmically transformed continuous variables to reduce the influence of outliers and z-score-standardized each column.

To tackle the irregular feature sampling rate, we applied piecewise cubic Hermite interpolation polynomials [11, 12]. After normalizing the timescale via interpolation, We filled remaining missing values - e.g. unobserved columns - with zeroes corresponding to the standardized empirical mean post z-score preprocessing.

2.4. Feature Selection

In our machine learning based algorithms, statistical features are selected by deriving various statistical metrics that represent different local and global aspects of the underlying signal. We compute sliding-window features for a variety of different window sizes to capture local and global descriptors of the ICU data. In total, we compute 18 features including moment statistics about the waveform distribution (mean, variance, skewness, kurtosis) as well as quantile information.

In our neural network methods, all 40 variables are exploited. The input is the processed data involving imputation and interpolation.

2.5. Classification

In this section, we summarize the details of a subset of methods we applied. We evaluated offline prediction via a variety of algorithms: Linear Least Squares
This dual-architecture has seen success in sequence processing tasks - e.g. video processing [17]. We provide an image of our architecture in Figure 3.

We evaluate the performance of our algorithms with a variety of metrics including accuracy, sensitivity, specificity of final and intermediate predictions. Furthermore, the challenge score and running time are considered.

3. Results and Analysis

We conclude that neural network, random forest, sparse quantile regression, naive Bayes, and the neighborhood
methods offer superior performance with respect to accuracy, sensitivity, and specificity. Table 1 displays the classification performance of each classifier. Random forests offer deceptively strong performance on average in comparison to the other algorithms, however the sensitivity is quite low. In contrast, sparse quantile regression outperforms other algorithms for sepsis detection and is robust to over fitting. Naive Bayes demonstrates balanced performance. Other classifiers show reduced capability for sepsis detection with ≤ 5% sensitivity.

### Challenge Scores

Our best challenge score is 0.227, which is based on our LSTM-CNN neural network. The best machine learning based approach achieved a challenge score of 0.056.

### Conclusion and Future Work

Our neural network is effective at learning complex patterns implicit in clinical data and associating them with sepsis, outperforming our baselines by a significant margin. Our future work, includes a more in-depth exploration of the developed techniques and integration of additional engineered features to this task. Additionally, we hope to experiment with methods which can implicitly handle irregularly sampled data and imbalanced class distribution.

### References


Address for correspondence:
Po-Ya Hsu
p8hau@eng.ucsd.edu