CMPUT 690 Project: Predicting the Onset of Sepsis

Humza S. Haider
Department of Computing Science
University of Alberta
Edmonton, AB T6E-4J5
hshaider@ualberta.ca

Abstract

Physionet and Computing in Cardiology have issued a challenge for the early prediction of sepsis in Intensive Care Unit patients. The data consists of 40,366 hospital patients of which only 2,932 ever have a septic event. Leveraging a wide array of tools including Missing Indicators (MIs), taking difference values between sequential time steps, and using sampling based methods and class weights we were able to overcome many of the challenges this competition presented. Our empirical results show our methods of XGBoost and Balanced Random Forests coupled with our data pre-processing steps highly outperform the baseline, rule based model for the prediction of sepsis. Experimental results suggest that even better results may be achieved through further ensembling of more complex models.

1 Introduction

Sepsis is a critical illness that can occur as the human body is responding to an infection (Singer et al., 2016). To combat an infection, chemicals are released by the immune system which can cause inflammation in the entire body – in extreme cases this can lead to organ failure and be fatal if left untreated. The Center for Disease Control reports[1] that 1.7 million adults in the United States develop sepsis every year, and 270,000 die as a result. Due to this large impact on healthcare, early detection of sepsis is a major research area facing clinicians and researchers today.

Early prediction of sepsis is key for increasing the rate of survival; studies have shown that early intervention with sepsis patients leads to a significant reduction in mortality (Rivers et al., 2001). Specifically, Kumar et al. (2006) suggested that each hour of delayed antimicrobial administration led to an average 4-8% decrease in the survival rate of sepsis patients. In an effort to increase the ability of clinicians to identify patients who will develop sepsis before it occurs, Physionet[2] and the Computing in Cardiology community have outlined a competition for the early prediction of the onset of sepsis[3]. This competition has been designed to replicate the real-life scenario of in hospital sepsis prediction, *i.e.* given the patient's *recent history* predict if they will experience sepsis in the next few hours. Here, *recent history* refers to the information given by vital signs, laboratory tests, and patient demographics over the time they have been in the hospital (typically for a few hours or days).

This type of classification problem poses a number of difficulties. The data is comprised of multivariate timeseries data of differing sequence lengths, *i.e* patients stay for differing amounts of time. Additionally, due the nature of real time medical classification the data will often contain many missing values. For example, patients may only have lab tests completed once every 12 hours, or may never have lab tests, leading to features with over 99% missing rates. On top of this, missing data is likely not Missing Completely At Random (MCAR), a common assumption when working with missing data. Since healthier patients may not need lab tests there will likely be a dependence between missing values and the patient's outcome, falsifying the MCAR assumption.

This work discusses our approach to dealing with the many difficulties this challenge presents. We first discuss and summarize the related work in Section 2, and follow with our methodology and empirical results in Sections 3 and 4. Future work and concluding remarks are made in Sections 5

[1] https://www.cdc.gov/sepsis/datareports/index.html

[2] https://physionet.org/

[3] https://physionet.org/challenge/2019/

and 6.

2 Related Work

In the last few years a number of papers have explored the use of a wide variety of machine learning methods for the early prediction of sepsis. Most notable has been the work of Calvert et al. (2016b), the developer of *Insight*, which has undergone numerous iterations and was recently included in a clinical trial to test its effectiveness in the early prediction of sepsis (Calvert et al., 2016b,a; Shimabukuro et al., 2017; Mao et al., 2018). The *Insight* model uses sliding windows on patient vital signs, calculating the feature means and feature differences between the window to incorporate the time dimension of the patient's features. To elaborate, if a sliding window of size five is used, the features will consist of the average of patients vital signs over 5 hours and the difference between the value of the vital sign at the current time and the value acquired 5 hours prior. After using the sliding window for feature construction, the constructed features are aggregated to create a risk score to classify patients as likely to experience sepsis in the next few hours and those who are unlikely to experience sepsis in the next few hours (Calvert et al., 2016b,a).

Newer iterations of the *Insight* model have used gradient tree boosting on these same vital signs for classification, however, instead of using a sliding window they include the both the current and previous time points as features in addition to the differences between the current and previous values to account for the temporal nature of the data (Mao et al., 2018).

Another recent work has explored the use of models from survival analysis – namely the Cox proportional hazards model with a Weibull base hazard function – for the early prediction of sepsis (Nemati et al., 2018). Similar to the work of Calvert et al. (2016b), this work made use of sliding windows and a survival model was used to predict if sepsis will occur in the next T hours, where they vary T throughout their experiments. Using this model they were able to outperform a number of baselines but no comparison to other machine learning models were made.

3 Methods

This section first describes the data and details the challenge presented by Physionet and Computing in Cardiology and then introduces the custom evaluation metrix used for the challenge. Following this definition, we describe the data preprocessing, feature construction, and models selected for this competition.

ICU	HR	…	Sepsis	Sepsis_Early
1	90	…	0	0
2	87	…	0	0
⋮	⋮	⋮	⋮	⋮
27	130	…	0	1
28	130	…	0	1
29	130	…	0	1
30	140	…	0	1
31	138	…	1	1
32	140	…	1	1

Table 1: An example data file for a single patient. Here ICU refers to the hours in the ICU and HR is the heart rate. Each patient has 8 vital signs, 26 possible laboratory tests, and 6 demographic variables. Note Sepsis_Early is the variable of interest as we wish to predict sepsis before it occurs. In total Sepsis_Early starts 12 hours prior to the onset of sepsis (here we only show 4 hours for brevity).

3.1 Challenge Definition and Data

The challenge issued by PhysioNet and Computing in Cardiology is designed to predict the early onset of sepsis in patients admitted to the Intensive Care Unit (ICU). In total, a population of 40,336 patients was supplied, with 2,932 having a sepsis event. Each patient observation is comprised of vital signs, laboratory tests, and demographics for every hour they stayed in the ICU. Across all patients there were an accumulative 1,552,210 hours of patient data, of which 27,916 hours were labeled as sepsis events.

As the goal of this challenge is early prediction, a sepsis event is defined as the patient experiencing sepsis as well as the twelve hours prior to experiencing sepsis. See Table 1 for an example of the data supplied for a single patient. Given this information, the outcome of a learned model is to predict if a patient will experience sepsis in the next twelve hours. This prediction is made for every hour a patient spends in the ICU. It is important to note that algorithms are not allowed to see future data, *e.g.* at 7 hours in the ICU, an algorithm is only allowed to use the information from hours 1-7 to predict if the patient will experience sepsis in the next twelve hours.

As part of the challenge, a custom evaluation function, referred to as the *Utility Score*, was created by Physionet to rank submissions. They define the Utility Score, $U(s,t)$, for each prediction of each patient (s) for each time interval (t):

$$U(s,t) = \begin{cases} U(s,t)_{TP} & \text{True Positive} \\ U(s,t)_{FN} & \text{False Negative} \\ U(s,t)_{FP} & \text{False Positive} \\ U(s,t)_{TN} & \text{True Negative} \end{cases}.$$

As seen above the Utility Score is made up of 4 sub-functions defined for the four types of classification. For example, a true positive would be predicting septic patient s is septic (or will become septic in the next twelve hours) at time t, whereas a false positive is predicting a non-septic patient s is septic when they are not nor will they become septic in the next twelve hours. As identifying septic patients is very crucial, an emphasis is applied to true positives and false negatives – see Figures 1 and 2 for an example plot of each of the four Utility functions. After calculation of the Utility Score it is then normalized such that is resides between 0 and 1 where a score of 0 represents a model which makes no predictions (all negatives) and 1 is the perfect model. Note a model can still be negative if it predicts many false positives, *e.g* a model which predicts all positives. For more details on this evaluation function please see the challenge website [4].

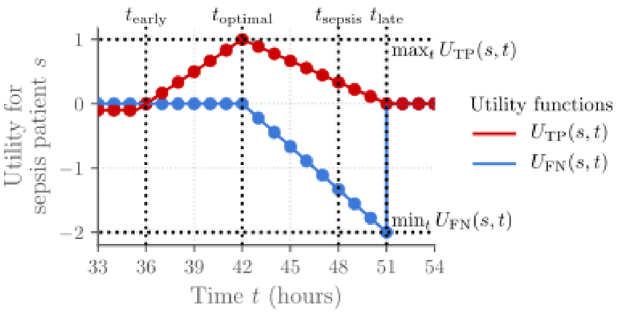

Figure 1: Utility function for true positives and false negatives. For a patient that starts experiencing sepsis at $t = 48$, the optimal prediction (6 hours prior to the onset of sepsis) counts for +1 towards the Utility Score. The largest penalty is incurred when a model fails to predict sepsis 3 hours after onset ($t = 51$), resulting in a penalty of -2.

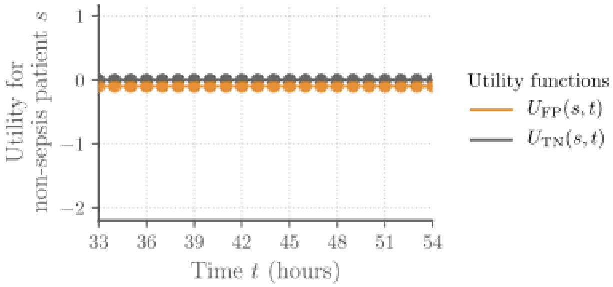

Figure 2: Utility function for false positives and true negatives. Since false positives only incur minor effort from hospital staff there is a small penalty of 0.05. True negatives give no penalty nor any points towards the Utility Score.

3.2 Data Pre-Processing

As previously mentioned, the data provided in this challenge contains large amounts of missing values – laboratory tests are typically collected once every twelve hours if at all. Excluding patient demographics, each patient feature is, on average, missing 80% of its values with a maximum missing rate of 99.8%. To cope with this we performed a forward-fill for each feature, *i.e.* the values for each feature were carried to future time intervals wherever it was applicable. Any other missing values (typically in the first few hours of stay in the ICU) were replaced by the mean of each feature.

In addition to the forward-fill method we also employ the use of Missing Indicators (MIs) introduced by Lipton et al. (2016). A MI uses a value of 1 to indicate that a feature's value was imputed (either by forward fill or using the mean imputation) and 0 if the value was actually observed. Lipton et al. (2016) showed that the use of these MIs significantly improved the performance of LSTMs and RNNs in a sequential medical classification task.

Due to the limited computational resources we chose to not apply neural networks for this work even though the processing of varied length sequential data is handled naturally by RNNs. Instead, we applied the methodology used by Mao et al. (2018) to incorporate patient history into the prediction. For each patient observation, Mao et al. (2018) obtained 3 values from each feature – the value at the current hour, the hour prior, and

[4] https://physionet.org/challenge/2019/

2 hours prior. In addition they took two difference values: the difference between the current hour and the hour prior and between the hour prior and the 2 hours prior. Thus, each original feature in a patients file accounted for 5 features post-transformations. We used this same methodology for our data, however, we only applied this to the vital signs which were collected every hour; since laboratory values were rarely collected there was no need to incorporate past values. After accounting for vital signs, difference values and past vital signs, laboratory values, demographics, and indicators there were a total of 107 features used for every line of every patient.

3.3 Models

We employed two primary methods for this challenge, extreme gradient boosted trees (XGBoost) and random forests. Very generally, XGBoost is an ensemble technique used to combine weak (shallow) decision trees iteratively such that each following classifier attempts to correct for mistakes made by previous trees (Chen and Guestrin, 2016). XGBoost is different than typical gradient boosting machines in that it uses a regularized model to help prevent overfitting and it was built to be scalable to very large data such as those provided in this challenge. While boosting naturally deals with imbalanced data by iteratively building trees on misclassified data, we additionally weight instances due to the class imbalance using a weight of $\frac{1{,}552{,}210}{27{,}916} \approx 55$ for positive instances.

For our random forest classifier, we use the method introduced by Breiman et al. (2004) known as the balanced random forests (BRF). Breiman et al. (2004) introduced BRFs to allow random forests to handle imbalanced data via an altered bootstrap sampling procedure. Instead of sampling a bootstrap size of N – the size of the dataset – a sample of $2N_1$ is chosen where N_1 is the number of instances of the minority class. Additionally, the minority class' sampling probability is adjusted such that they are sampled evenly with the majority class. For BRFs we also use the variable selection strategy of OBrien and Ishwaran (2019), who showed significant improvement in random forests under imbalanced data. Features are selected by their variable importance as measured by the Ishwaran-Kogalur importance using G-mean prediction error (Ishwaran and Lu, 2019). If variables were found non-significant at the 5% level they were discarded (significance as calculated by methods described in Ishwaran and Lu (2019)).

Criteria	**Threshold**
Body Temp (C)	<36 or >38
Heart Rate	>90
White Blood Cell Count	<4 or >12
Respiratory rate	>20

Table 2: The SIRS criteria and thresholds commonly used to identify sepsis.

4 Experiments

Since their is a significant amount of data we chose to evaluate on a single train/test split as opposed to cross validation measures. The data was split into 80% train, 20% test such that individual patients were entirely contained either within the training set or the test set, *e.g.* Patient 1 could not have lines in the training set and the test set. Ideally we would be able to compare against other models submitted to the challenge, however, at the current time the leaderboard for the competition has not yet been made public. The competition's preliminary phase ends on April 15, 2019 at which point the leaderboard will be made available. For this reason we have showed our results on a train/test split as opposed to comparing against the competition test set.

As a baseline model, the early diagnostic criteria, Systematic Inflammatory Response Syndrome (SIRS) method was used (Rangel-Frausto et al., 1995). SIRS uses four criteria (see Table 2) to diagnosis sepsis and is commonly used as a baseline model in early sepsis prediction (Calvert et al., 2016b,a; Mao et al., 2018). While the SIRS score ranges from zero to four, a value of 2 is indicative of sepsis and is commonly used as the threshold for diagnosis and as such we use SIRS ≥ 2 to delineate a positive sepsis prediction.

XGBoost was trained for 1500 iterations with positive class weights of 55, and with a subsampling proportion of 0.5 for computational speed and to prevent overfitting. BRFs were trained using 3000 trees as suggested by Ishwaran and Kogalur (2019). In addition to the BRF models (both with and without feature selection), we also consider the average of the predictions between the XGBoost model and the BRF model.

Model	Utility Score	AUROC
SIRS	0.062	0.617
BRF - No FS	0.385	0.834
BRF - FS	0.384	0.833
XGBoost	0.411	0.834
Avg. Prediction	**0.414**	**0.840**

Table 3: The Utility Score and AUROC for the primary models considered (higher scores are better for both metrics). For the BRF models, FS stands feature selection and **bolded** values indicate the best performing model.

The Utility Score defined in Section 3 and the area under the receiver operating characteristic curve (AUROC) are used as our evaluation metrics and are reported in Section 4.1. An ablation analysis studying the usefulness of MIs and the difference features (DFs) are presented in Section 4.2

4.1 Results

Table 3 shows the results of the models tested on both the Utility Score and the AUROC. Of note is that all models greatly surpasses the baseline SIRS model. While SIRS scored very poorly in terms of the Utility Score and the AUROC, it was similar to results found using comparable datasets in prior works (Mao et al., 2018; Gupta et al., 2018). It appeared that SIRS had an extremely high false positive rate, lowering its overall score.

Between the BRF and the XGBoost models, XGBoost outperformed BRF (both with feature selection and no feature selection) on the Utility Score but were equivalent for the AUROC. We found feature selection was not helpful for the BRF model and so was not tested with the XGBoost model nor considered in the ablation analysis in the following section.

While XGBoost outperformed the BRF models, when the prediction probabilities between BRF (no feature selection) and the XGBoost model were averaged, they slightly outperformed all the other models. This suggests that an ensemble of multiple model types may be beneficial for future performance gains.

4.2 Ablation Analysis

In order to consider the impact that the MI and difference features (DFs) have on the analysis we reran the analysis without using the MI features and without the DFs separately. The results from this analysis are presented in Table 4.

Model	Utility Score	AUROC
BRF - No FS	0.385	0.834
BRF - No FS/MI	0.368	0.828
BRF - No FS/DF	0.379	0.832
XGBoost	0.411	0.834
XGBoost - No MI	0.404	0.832
XGBoost - No DF	0.413	0.834

Table 4: Ablation analysis of the Missing Indicator features introduced by Lipton et al. (2016) and the difference features (DFs) used by Mao et al. (2018). Both BRFs and XGBoost with and without MIs are considered here. **Bolded** values indicate the best performing models within each subgroup (BRF and XGBoost). Values for XGBoost and BRF - No FS are repeated from Table 3 for comparability.

While not having a profound impact, it is clear that including the MI features are useful to both the XGBoost and BRF models. BRF in particular appeared to be impacted by the inclusion of MI features as the Utility score rose by 0.077 (more than the total score achieved by SIRS), but AUROC only increases slightly from 0.828 to 0.834.

Interestingly, scores of XGBoost actually *improved* when the difference features were removed. While only a marginal difference, the Utility Score for XGBoost rose by 0.002 and the AUROC remained constant. For BRF the Utility Score and AUROC dropped a very small amount, 0.006 and 0.002 respectively. This is suggestive that the DFs are not actually beneficial and may instead be acting as noisy features. Rather, it appears models can rely on the realized values of vital signs for the early prediction of sepsis without additional feature construction.

5 Future Work

The official competition deadline is August 25, 2019 so there is lots of opportunity to make improvements to the model. Prior work has shown that using heuristic tables similar to those used by the rule based systems used in practice (*e.g.* SIRS) can greatly improve the performance of different machine learning models for the early prediction of sepsis (Calvert et al., 2016b,a). Using such methods may improve our own XGBoost and BRF models as well.

Our basic ensemble of XGBoost and the BRF model suggests that performance may improve as more models are ensembled together. This opens the door to using more complex models such as neural networks for prediction and can be further optimized through ensembling.

6 Conclusion

The Physionet challenge for the early prediction of sepsis posed many challenges including varying length multivariate sequential data, a large class imbalance, and an inordinate amount of missing data. By combining various techniques for data pre-processing and feature construction we were able to resolve these issues and build an XGBoost model and an BRF model which substantially outperformed the rule based SIRS baseline model. Future gains were observed through a simple averaging of model results suggesting that more complex ensembling methods of different types of models may achieve even higher performance.

An ablation analysis suggested that while MI features were beneficial for prediction, the DFs used by Mao et al. (2018) were not beneficial and could actually be detrimental to predictive ability of the models considered. In the future we plan to utilize heuristic tables and ensemble additional types of models to raise the performance of our system in the Physionet 2019 challenge.

References

Leo Breiman, Chao Chen, and Andy Liaw. 2004. Using random forest to learn imbalanced data. *J. of Machine Learning Research* 666.

Jacob Calvert, Thomas Desautels, Uli Chettipally, Christopher Barton, Jana Hoffman, Melissa Jay, Qingqing Mao, Hamid Mohamadlou, and Ritankar Das. 2016a. High-performance detection and early prediction of septic shock for alcohol-use disorder patients. *Annals of medicine and surgery* 8:50–55.

Jacob S Calvert, Daniel A Price, Uli K Chettipally, Christopher W Barton, Mitchell D Feldman, Jana L Hoffman, Melissa Jay, and Ritankar Das. 2016b. A computational approach to early sepsis detection. *Computers in biology and medicine* 74:69–73.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, pages 785–794.

Akash Gupta, Tieming Liu, Scott Shepherd, and William Paiva. 2018. Using statistical and machine learning methods to evaluate the prognostic accuracy of sirs and qsofa. *Healthcare informatics research* 24(2):139–147.

H. Ishwaran and U.B. Kogalur. 2019. *Random Forests for Survival, Regression, and Classification (RF-SRC)*. R package version 2.8.0. https://cran.r-project.org/package=randomForestSRC.

Hemant Ishwaran and Min Lu. 2019. Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in medicine* 38(4):558–582.

Anand Kumar, Daniel Roberts, Kenneth E Wood, Bruce Light, Joseph E Parrillo, Satendra Sharma, Robert Suppes, Daniel Feinstein, Sergio Zanotti, Leo Taiberg, et al. 2006. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Critical care medicine* 34(6):1589–1596.

Zachary C Lipton, David C Kale, and Randall Wetzel. 2016. Modeling missing data in clinical time series with rnns. *Machine Learning for Healthcare* .

Qingqing Mao, Melissa Jay, Jana L Hoffman, Jacob Calvert, Christopher Barton, David Shimabukuro, Lisa Shieh, Uli Chettipally, Grant Fletcher, Yaniv Kerem, et al. 2018. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and icu. *BMJ open* 8(1):e017833.

Shamim Nemati, Andre Holder, Fereshteh Razmi, Matthew D Stanley, Gari D Clifford, and Timothy G Buchman. 2018. An interpretable machine learning model for accurate prediction of sepsis in the icu. *Critical care medicine* 46(4):547–553.

Robert OBrien and Hemant Ishwaran. 2019. A random forests quantile classifier for class imbalanced data. *Pattern recognition* 90:232–249.

M Sigfrido Rangel-Frausto, Didier Pittet, Michele Costigan, Taekyu Hwang, Charles S Davis, and Richard P Wenzel. 1995. The natural history of the systemic inflammatory response syndrome (sirs): a prospective study. *Jama* 273(2):117–123.

Emanuel Rivers, Bryant Nguyen, Suzanne Havstad, Julie Ressler, Alexandria Muzzin, Bernhard Knoblich, Edward Peterson, and Michael Tomlanovich. 2001. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *New England Journal of Medicine* 345(19):1368–1377.

David W Shimabukuro, Christopher W Barton, Mitchell D Feldman, Samson J Mataraso, and Ritankar Das. 2017. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ open respiratory research* 4(1):e000234.

Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. 2016. The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama* 315(8):801–810.