Early Prediction of Sepsis Using Gradient Boosting Decision Trees with Optimal Sample Weighting

Ibrahim Hammoud[1*], IV Ramakrishnan[1*], Mark Henry[2*]

[1] Department of Computer Science
[2] Department of Emergency Medicine
[*] Stony Brook University, Stony Brook, US

Abstract

In this work, we describe our early sepsis prediction model for the PhysioNet/Computing in Cardiology Challenge 2019. We prove that maximizing a general family of utility functions (of which the challenge utility function is a special case) is equivalent to minimizing a weighted 0-1 loss. We then utilize this fact to train an ensemble of gradient boosting decision trees using a weighted binary cross-entropy loss.

Our model takes the time-series nature of the data into account by using a fixed size window of all measurements within the last 20 hours as a feature vector. Data were imputed in a way that gives the same information to the model as present to healthcare professionals in real-time. We tune the model hyper-parameters using 5-fold cross-validation. The model performance was measured on each evaluation set using the threshold that gives the maximum utility on the training set. Our best model achieves an official normalized utility score of 0.332 on the final full test set of the challenge (Team name: SBU, rank: $6^{th}/78$).

1. Introduction

Sepsis is a critical medical condition caused by the body's response to infection [1, 2]. Some studies estimate that nearly half of hospital mortalities occur among patients with sepsis, making it one of the leading causes of deaths in hospitals [3].

Sepsis accounts for more than $20 billion of total US hospital costs [4]. Moreover, the incidence rate of sepsis is increasing [5]. Early prediction of sepsis can help in lowering these costs in addition to aiding early intervention and improving healthcare outcomes in hospitals [6, 7].

The PhysioNet/Computing in Cardiology Challenge 2019 is organized for early prediction of sepsis using automated approaches [8]. In this paper, we present our work for this challenge based on our best submission using gradient boosting decision trees.

2. Methods

2.1. Terminology

Denote by $X = \{x_i\}_{1 \leq i \leq n}$ a set of datapoints and by $y \in \{-1, +1\}^n$ the set containing their labels. We associate with each x_i a vector c_i containing context information that might not be available at testing time. In the setting of sepsis prediction, we define $c_i = (t_i, o_i)$ where t_i is the time until the sepsis event (if it exits) and o_i is the outcome type (developed sepsis or not) for the i^{th} datapoint corresponding to a certain patient.

We denote by $U_{x_i,c_i}(\hat{y})$ the utility of predicting $\hat{y} \in \{-1, +1\}$ for a given datapoint x_i with its context vector c_i. This terminology is a short hand notation to summarize the utility function in [8] whose output depends on time, prediction class and whether the patient eventually developed sepsis or not.

2.2. Imputation

In our work, we follow an imputation scheme that mimics the information present to healthcare professionals in real-time. At $t = 0$ all missing features are replaced with the training set mean. At $t > 0$, for every feature, if the feature is missing at the current time, we replace the feature with the last available value. We also append for every feature the number of hours it has been missing as a new feature.

Another advantage of this imputation scheme is that it is convenient for the learning process of decision tree models. It allows them to learn the relationship between the relevance of certain features (or a combination of them) based on how long each of them has been missing.

2.3. Sample weighting

The utility function given in the challenge varies with time, patient condition and prediction output. Using a regular binary labeling of the data-points without an appropri-

ISSN: 2325-887X DOI: 10.22489/CinC.2019.459

ate weighting is not ideal since data-points within the same class impact the total utility with different magnitudes.

More precisely, one can see that incorrectly predicting a single point results in decreasing the total utility by the difference between its correct prediction utility and its incorrect prediction utility. Using this observation, we arrive at Theorem 1 below which is true for the general case and not only in the context of the challenge utility function.

Theorem 1. *Maximizing the classification utility of a dataset X according to a utility function $U_{x_i,c_i}(\hat{y})$ is equivalent to minimizing a weighted 0-1 loss.*

$$\arg\max_{h\in H}\sum_{i=1}^{|X|} U_{x_i,c_i}(\hat{y}_i) = \arg\min_{h\in H}\sum_{i=1}^{|X|}[\hat{y}_i \neq sign(\gamma_i)] \times |\gamma_i|$$

$$where\ \ \gamma_i = U_{x_i,c_i}(+1) - U_{x_i,c_i}(-1).$$

Proof. Given a hypothesis space H, let $\hat{y}_i$ be the predicted class according to classifier $h \in H$. Moreover, in what follows we denote by $\hat{y}_i^{'}$ the opposite class of $\hat{y}_i$. Given this, we can see that:

$$\begin{aligned}
\arg\max_{h\in H}\sum_{i=1}^{|X|} U_{x_i,c_i}(\hat{y}_i) &= \arg\min_{h\in H}\sum_{i=1}^{|X|} U_{x_i,c_i}(\hat{y}_i^{'}) \\
&= \arg\max_{h\in H} -\sum_{i=1}^{|X|} U_{x_i,c_i}(\hat{y}_i^{'}) \\
&= \arg\max_{h\in H}\sum_{i=1}^{|X|} U_{x_i,c_i}(\hat{y}_i) - \sum_{i=1}^{|X|} U_{x_i,c_i}(\hat{y}_i^{'}) \\
&= \arg\max_{h\in H}\sum_{i=1}^{|X|}(-1)^{[\hat{y}_i\neq sign(\gamma_i)]}|\gamma_i| \\
&= \arg\min_{h\in H}\sum_{i=1}^{|X|}\Big(\big(1-(-1)^{[\hat{y}_i\neq sign(\gamma_i)]}\big)|\gamma_i|\Big) \\
&= \arg\min_{h\in H}\sum_{i=1}^{|X|}[\hat{y}_i \neq sign(\gamma_i)] \times |\gamma_i|.
\end{aligned}$$

□

Corollary 1. *Maximizing the challenge utility function is equivalent to minimizing a weighted 0-1 loss such that:*

$$\begin{aligned}
&X = \{x_i\}_{1\leq i\leq n} \\
&C = \{c_i = (t_i, o_i)\}_{1\leq i\leq n} \\
&y = \{y_i = sign(\gamma_i)\}_{1\leq i\leq n} \\
&w = \{w_i = |\gamma_i|\}_{1\leq i\leq n} \\
&\gamma_i = U_{x_i,c_i}(+1) - U_{x_i,c_i}(-1).
\end{aligned}$$

Given this, we use a weighted binary cross-entropy loss to train our model as it is commonly used in classification tasks to approximate the 0-1 loss:

$$L(y, y^*) = \frac{1}{2n}\sum_{i=1}^{n} -w_i\Big((1+y_i)\log(y_i^*) + (1-y_i)\log(1-y_i^*))\Big)$$

$$where\ \ y_i^* = P(y = 1|x).$$

2.4. Model

To build an early sepsis prediction model, we trained an ensemble of gradient boosting decision trees (GDBT) [9]. The model was trained and evaluated on the combined dataset from both hospitals A and B containing 40336 patients out of which 2932 had a positive sepsis label. The model was implemented using the lightGBM library (version 2.2.3) [10]. The max tree depth was fixed at $d = 7$ and the number of learners was fixed at 100. We then tuned the regularization strength parameters for L_1 and L_2 regularization in the ranges (0, 50) and (0, 500) respectively. The optimal values were selected based on 5-fold cross-validation.

To accommodate for the sequential nature of the data, for each patient at time t, we augment the datapoints of the last 20 hours into one vector. If $t < 20$, we augment zero vectors. Ideally taking the 60 last hours would have been a better choice as the majority of stays (more than 97%) are below 60 hours. However, due to computational limitations, we fixed the number at 20 hours.

The label and weight for each data point were selected according to Corollary 1. The model was trained to minimize the weighted binary cross-entropy discussed earlier.

2.5. Score calibration

After fixing the model parameters, and for each fold, a linear time algorithm is used to try all possible thresholds on the training set. The best threshold resulting in maximum utility on the training set is used to evaluate the utility on the evaluation set. The model resulting in the best utility score on the evaluation set across the 5 folds is then submitted for the final testing.

3. Results

Table 1 summarizes our official final challenge results. We used 4 out of our 10 allowed submissions. Our best model achieves a utility score of 0.332 on the full test set. Our final team ranking was 6th.

Team Name	SBU
Final test set normalized utility score	0.332
Rank	$6^{th}/78$

Table 1: Official final challenge results

In what follows, we analyze our model output scores and how the given utility function impacts the overall model performance. More generally, we are interested in how optimizing for the challenge utility function can impact real-time prediction and decision making. Under this setting, the time of prediction for positive patients is the first time their scores go above a fixed threshold. Negative patients are considered false positives if at any point in time their scores go above the same fixed threshold. The results below reflect the model performance on the evaluation set using one of the 5 folds (80% training and 20% evaluation on the combined set of hospital A and hospital B patients).

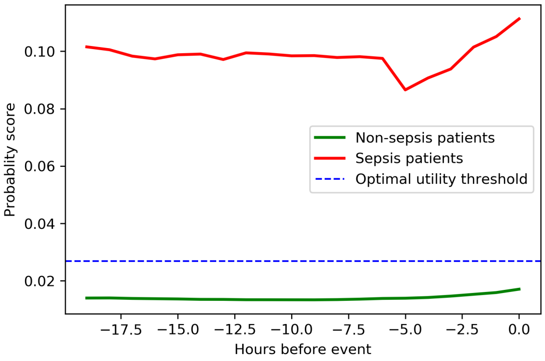

Figure 1: Average model output probability scores for sepsis and non-sepsis patients. For sepsis patients, the time on the x-axis is relative to the sepsis event time. For non-sepsis patients, the time is calculated relative to the time of their last recorded measurement in the dataset. We note that there is a general trend of increase in the average scores for sepsis patients between t_{sepsis} - 5 and t_{sepsis}. This is in alignment with the increasing weight of points in that interval according to the utility function.

As shown in Figure 1, there is a general trend of increase in the average score for positive patients as we approach the sepsis event. We also see that the optimal utility threshold separates the average scores of positive and negative patients. However, even with such a wide gap between the average scores, we still observe a high false positive rate at the point of maximum utility (Figure 2). This is mainly because of the high variance of the score distribution of positive patients as shown in Figure 3.

In Figure 4, we can see that the median detection time among all patients in the evaluation set is always above 25 hours. This is considered a very early prediction for sepsis (penalized by the challenge utility function). The reason the detection time is not going lower can be illustrated in Figure 5. Out of 407 patients who had sepsis and had records 15 hours before the sepsis event, 283 have higher scores 15 hours before the sepsis event than at the sepsis event. Because of this, no matter what threshold we set, for most of sepsis patients, an alarm would be raised much before the onset of sepsis.

From an optimization perspective, this is an artifact of the utility function described in [8]. The weight of the point at the sepsis event is 33.33 times that of any point 12 hours before it. The model might choose a slight increase in the score of a sepsis event point at the expense of a substantial increase in the probability scores of multiple points much before the event. A similar behavior can be seen for points ∈ (t_{sepsis} - 6, t_{sepsis} + 3) (which have high weights). The collective effect of this results in having a high number of early points with high probability scores, causing positive patients to raise alarms very early on.

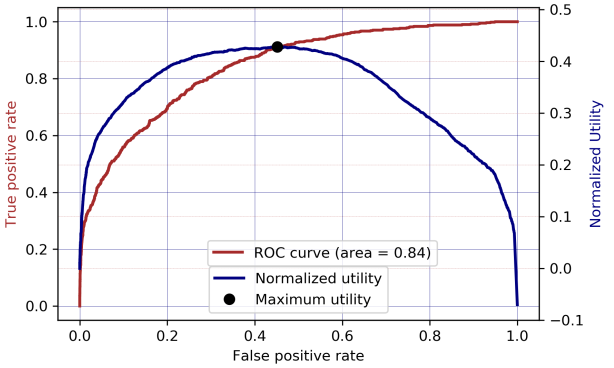

Figure 2: ROC curve and normalized utility as a function of false positive rate plot. At the point of maximum utility, the model has a high false positive rate of 0.451. ROC curve was computed at patient level.

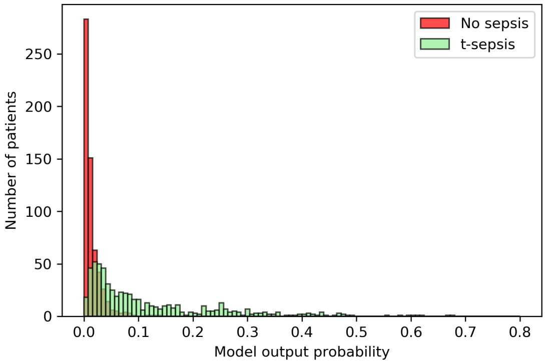

Figure 3: Comparison between the model output probability scores for sepsis and non-sepsis patients. The green histogram shows the model output probability for 605 sepsis patients at the sepsis event time. For non-sepsis patients, we first selected 605 of them randomly and then chosen one random point from each patient.

4. Conclusion

The results in this paper show that gradient boosting regression trees perform relatively well for early prediction of sepsis. In this dataset, data-points were bucketed into hours. On a different dataset with real-time signals measured with high frequency such as heart rate, bucketing would cause significant loss of information. This would make it challenging to design fixed-size feature vectors that incorporate short and long term information. In such a scenario, sequence models such as LSTMs are more suit-

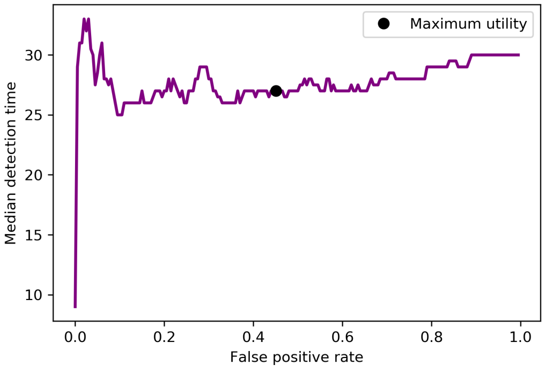

Figure 4: Median detection time as a function of false positive rate. The plot shows at each false positive rate, the median detection time for positive patients who were correctly predicted by the model. Note that the detection time is always above 25 hours. At the point of maximum utility (0.451 FPR), the median detection time was 27 hours. For low FPR values, the number of correctly predicted patients is small, which explains the higher variance in that region.

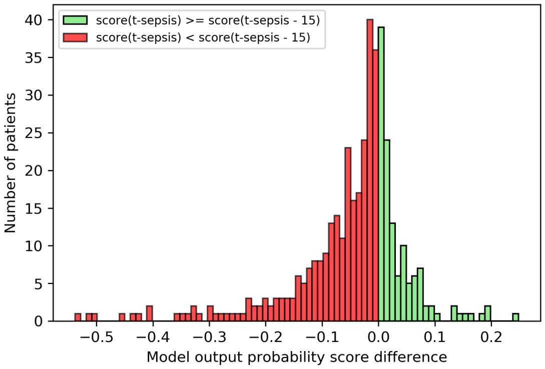

Figure 5: Distribution of the difference between the model output probability scores at the sepsis event and 15 hours before the event for 407 patients. There are 283 patients who had higher scores 15 hours before the sepsis event than at the sepsis event. A similar behavior can be noticed at multiple different time points too. This results in a very early prediction that doesn't change much as we increase the decision threshold explaining the shape of the graph in Figure 4.

able. They learn to capture those patterns at training time given the full original time series data.

Motivated by the impacts of the utility function on the model output scores shown in the results section, we would like to explore more metrics, scoring functions and models for early prediction tasks in future work. Particularly, we are interested in developing methods that better optimize for the real-time prediction setting discussed in this paper.

Acknowledgments

Research partially supported by the Emergency Department in the Renaissance School of Medicine at Stony Brook University and NIH award HD097188-01.

References

[1] Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, Rubenfeld G, Kahn JM, Shankar-Hari M, Singer M, et al. Assessment of clinical criteria for sepsis: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). Journal of the American Medical Association 2016;315(8):762–774.

[2] Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche JD, Coopersmith CM, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). Journal of the American Medical Association 2016;315(8):801–810.

[3] Liu V, Escobar GJ, Greene JD, Soule J, Whippy A, Angus DC, Iwashyna TJ. Hospital Deaths in Patients With Sepsis From 2 Independent CohortsHospital Deaths in Patients With SepsisLetters. JAMA 07 2014;312(1):90–92. ISSN 0098-7484. URL https://doi.org/10.1001/jama.2014.5804.

[4] Torio CM AR. National inpatient hospital costs: The most expensive conditions by payer, 2011: Statistical brief 160., 2013.

[5] Iwashyna TJ, Cooke CR, Wunsch H, Kahn JM. Population burden of long-term survivorship after severe sepsis in older americans. Journal of the American Geriatrics Society may 2012;60(6):1070–1077.

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

[7] Seymour CW, Gesten F, Prescott HC, Friedrich ME, Iwashyna TJ, Phillips GS, Lemeshow S, Osborn T, Terry KM, Levy MM. Time to treatment and mortality during mandated emergency care for sepsis. New England Journal of Medicine 2017;376(23):2235–2244.

[8] 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.

[9] Friedman JH. Greedy function approximation: A gradient boosting machine. The Annals of Statistics 2001;29(5):1189–1232. ISSN 00905364. URL http://www.jstor.org/stable/2699986.

[10] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. Lightgbm: A highly efficient gradient boosting decision tree. In Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds.), Advances in Neural Information Processing Systems 30. Curran Associates, Inc., 2017; 3146–3154.

Address for correspondence:

Ibrahim Hammoud
350 Circle Rd, Stony Brook, NY, 11790, US
ihammoud@cs.stonybrook.edu