A Personalized Boosting Screening Method and Its Application to Sepsis
Chen Feng\textsuperscript{a}, Paul Griffin\textsuperscript{b}, Shravan Kethireddy\textsuperscript{c}, and Yajun Mei\textsuperscript{a}

\textsuperscript{a}School of Industrial & Systems Engineering, Georgia Tech, Atlanta, GA; \textsuperscript{b}Regenstrief Center for Healthcare Engineering, Purdue University, West Lafayette, IN; \textsuperscript{c}Critical Care Medicine, Northeast Georgia Medical Center, Gainesville, GA

Motivation and Objective

- The natural mortality rate for sepsis is between 25% and 50%.
- Nearly half of patients who die in hospitals are septic.
- In 2016, a task force committee recommended screening for sepsis by quick Sepsis-related Organ Failure Assessment (qSOFA), which uses the constant thresholds in decision making.

![SEPSIS BEDSIDE CRITERIA](image)

- **Our Objective:** Develop a personalized sepsis screening method that depends on a patient’s baseline characteristics such as age, sex, admission location, etc.

Problem Formulation

- **Data:** \((Y_i, X_i, u_i), \) for \(i = 1, \ldots, N, \) where \(Y_i \in \{-1, 1\} \) is the binary outcome, \(X_i \) is the biomarker (e.g., blood pressure, respiratory rate, etc.), and \(u_i \in R^{p_0} \) are baseline characteristics.

- **Classification Rule:** Predict \(Y_i \) by

\[
Y_i = \begin{cases} 1, & \text{if } X_i \geq c(u_i) \\ -1, & \text{otherwise} \end{cases}
= \text{sign}(X_i - c(u_i)).
\]

Here we assume
\[
c(u_i) = u_i^T \beta,
\]
for some unknown parameters \( \beta = (\beta_0, \beta_1, \ldots, \beta_{p_0})^T. \)

- **Question:** Estimate \( \beta \) so as to minimize misclassification rate.
  - The function sign(\( f \)) is not continuous.
  - The \((0-1)\) loss function \( I(Y \neq \text{sign}(f)) \) is non-smoothing.
  - The consequences of misclassifying sepsis and non-sepsis patients are different.

Existing Methods

qSOFA considers the constant threshold, i.e., \( c(u_i) \equiv c. \)

Several approaches to find suitable constant thresholds:
- **minP Approach:** Maximizing the standard chi-square statistic
- **Youden Index:** Maximizing the sum of sensitivity and specificity
- **Closest-to-(0,1) Criterion:** The "optimal" threshold is defined as the point on the ROC curve closest to \((0,1)\)

Our Proposed Method: Personalized Threshold

- **Key ideas in our proposed method**
  - We define the threshold \( c_i(u_i) \) as a function of the individual subjects’ characteristics \( u_i. \)
  - We borrow the idea of boosting to replace the \((0-1)\) loss function \( I(Y \neq \text{sign}(f)) \) by the exponential loss \( \exp(-Yf) \)
  - Introducing two different weights, \( w_i \) and \( w_{-i} \), depending on whether \( Y_i = +1 \) or \( -1, \) in order to take into account the different consequences of misclassification.

Parameter Estimation as Optimization Problem

We propose to estimate the \((q+1)\)-dimensional parameter \( \beta \) by minimizing the training error under the weighted exponential loss function:

\[
J(\beta) = \frac{1}{N} \sum_{i=1}^{N} \left[ w_i \cdot \exp\left(-Y_i f_i\right) \cdot I(Y_i = 1) + w_{-i} \cdot \exp\left(Y_i f_i\right) \cdot I(Y_i = -1) \right],
\]

where \( f_i = f(X_i, u_i) = X_i - c_i(u_i) \) and \( w_i = \left(w_i + (Y_i + 1) + w_{-i} \cdot (1 - Y_i)\right)/2. \)

Computational algorithm: Gradient descent

**Algorithm 1:** Gradient Descent Using Weighted Exponential Loss

**Require:** \( Y, X, U, N, w_i > 0, w_{-i} > 0, \) \( n, T \)

1. Initialization: \( \beta_i = 0 \) \( \forall i \in \{0, 1, 2, \ldots, q\}, \) \( W = (w_i Y_i + 1) + w_{-i} (1 - Y_i))/2 \)
2. for all \( i = 1, 2, \ldots, T \) do
3. \( f_i = X_i - c_i(u_i) \)
4. \( f = X_i - c_i(u_i) \)
5. \( J = \frac{1}{N} L \)
6. \( d = \frac{1}{N} W 
\) \( \exp(-Y f) \)
7. \( \beta = \beta - \alpha \frac{d}{\partial \beta} \)
8. end for
9. \( \beta = \beta - \alpha \frac{d}{\partial \beta} \)

**Proposition:** The objective function \( J(\beta) \) is convex with respect to \( \beta, \) and thus the gradient descent algorithm converges to the global optimum if the learning rate \( \alpha \) is small enough and the optimization steps \( T \) is long enough.

Data Set

Medical Information Mart for Intensive Care III (MIMIC-III) database (version 1.4)

**Study Population:** 3,771 sepsis patients \((Y = 1); 4,000\) non-sepsis patients \((Y = -1)\)

**qSOFA variables \((X)\):** respiratory rate (RR), systolic blood pressure (sysBP), and GCS scores (we keep the constant cutoff 15 for GCS score, since GCS score is a discrete variable)

**Baseline characteristics \((u_i)\):** age, gender, admission location, admission type, ethnicity, insurance, and marital status

Application to Sepsis Screening

Comparison to qSOFA thresholds and existing methods

The parameters \( T = 30000, \alpha = 0.001, w_i = 1, \) and \( w_{-i} = 1 \) were selected based on a grid search to maximize the prediction accuracy.

<table>
<thead>
<tr>
<th>Method</th>
<th>Overall Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personalized qSOFA</td>
<td>0.7093</td>
<td>0.6668</td>
<td>0.6905</td>
</tr>
<tr>
<td>Closest-to-(0,1) criterion</td>
<td>0.7093</td>
<td>0.6668</td>
<td>0.6905</td>
</tr>
</tbody>
</table>

**Interpretation and Implementation of Personalized qSOFA**

A) The estimated personalized threshold for RR against age.

B) Compared to qSOFA for two specific patients (Left: Screening for non-sepsis patient. Right: Screening for sepsis patient)

C) Compared to Machine Learning Techniques

**Table:**

<table>
<thead>
<tr>
<th>Method</th>
<th>Overall Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personalized qSOFA</td>
<td>0.7093</td>
<td>0.6668</td>
<td>0.6905</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.7489</td>
<td>0.7742</td>
<td>0.7548</td>
</tr>
<tr>
<td>AdaBoosting</td>
<td>0.7359</td>
<td>0.7359</td>
<td>0.7359</td>
</tr>
</tbody>
</table>

**Contact Information**

Chen Feng, email: cfeng@gatech.edu