EARLY PREDICTION OF SEPSIS FROM CLINICAL DATA

Po-Ya Hsu
Department of Computer Science and Engineering
University of California San Diego
La Jolla, CA 92093
p8hsu@eng.ucsd.edu

Chester Holtz
Department of Computer Science and Engineering
University of California San Diego
La Jolla, CA 92093
chholtz@eng.ucsd.edu

April 15, 2019

ABSTRACT

Aims: In this study, we intend to explore the efficacy of modern machine learning methods for the task of modeling sepsis progression.
Methods: For our *initial investigation*, we developed a novel imputation and feature selection scheme based on signal processing technology and our medical expertise. To address data imbalance issues, we experimented with various approaches including sample weighting, re-sampling via balanced bagging and combinations of patient co-variates. We performed imputation with interpolation via cubic splines and utilized rolling window statistics for our features. We empirically chose the window size to be three hours. The feature matrix was composed of the maximum of summation, variation, and maxima, and the minimum of the minima of the rolling interpolated data. We compared the performance of several baseline classification algorithms including Support Vector Machine (SVM), hidden Markov models (HMM), naive Bayes, logistic regression (LR), gradient boosting (XGboost), random forest, sparse quantile regression, and Gaussian process (GP) classification. We performed cross validation to tune the optimal windows size and patient co-variates of interest. In our cross validation, the training and testing data sizes were 75% and 25%, and we adopted a binary classification framework.
Results: We conclude that random forest, sparse quantile regression, and naive Bayes classifiers 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 (XGboost, SVM, GP, SGD, LR, HMM) show low capability in sepsis detection with $\leq 5\%$ sensitivity.
Future work: For future work, we plan to explore incorporation of additional engineered features and the application of neural networks to this task.

Table 1: Performance of Sepsis Detection based on Processed Vital Signs

Classifier	Accuracy	Sensitivity	Specificity
Naive Bayes	84%	25%	90%
Random Forest	91%	9%	99%
Sparse Quantile Regression	60%	66%	58%
XGboost, SVM, GP, SGD, LR, HMM	$\approx 90\%$	$\leq 5\%$	$\approx 99\%$