An Automated Algorithm for Early Prediction of Sepsis

Reza Firoozabadi*, Saeed Babaeizadeh
Advanced Algorithm Research Center, Philips Healthcare, Andover, MA, USA

Method: We used the extended database including two datasets: dataset A with 20,336 and dataset B with 20,000 subjects. Each subject has a table of 40 time-dependent features and a sepsis onset label based on the Sepsis-3 definition over time. We imputed the missing measurements with their latest values if available, otherwise with their population median. Features were standardized by their population mean and standard deviation across all subjects. To avoid overfitting, for a sepsis subject we selected only the features one hour before the sepsis onset and for non-sepsis patients the measurements at a time near the end of interval were selected.
An ensemble of bagged decision trees was defined to classify the patient measurements at each time. The classifier was trained on dataset A where it was split into 80% training and 20% evaluation subsets. Validation was done using the whole dataset B. Important features were evaluated using this classifier and 15 most-important features were selected, of which ICU length of stay and heart rate were the most important features. The least important features were Unit1, Unit2 and Bilirubin_direct. Other important features were BUN, WBC, BaseExcess, and Calcium. Dataset A.

Results: The scores for the submitted unofficial entries were not available at the time of preparation of this abstract. However, our utility score computations using the provided code show that the evaluation subset of dataset A has 74% area under ROC curve, 91% accuracy, and a utility score of 0.32. Utility scores computations for dataset B result in 73% AUC, 93% accuracy and a utility score of 0.23. We intend to compare and improve our algorithm by designing a recurrent neural network time-series prediction algorithm (an LSTM network).