Machine Learning and Statistical Techniques to Predict Sepsis: Unifying Previous Work

Nathaniel D. Bastian, Paul M. Griffin, Shravan Kethiredy, Joseph J. Klobusicky, Alexander Murph, Alexander Robinson, Nathan C. Ryan

Introduction

Sepsis is a syndrome of dysregulated inflammation caused by infection. In 2010 it was that 11th leading cause of death in the United States [1] and, in 2011, it was the single most expensive diagnosis treated in hospitals in the United States [2]. Septic shock is a subset of sepsis in which the complications induced by the presence of sepsis lead to an increase in the mortality of the patient.

Methods

Finding a Septic Shock Diagnosis

Following a method developed by Ho et al. [3] for determining Septic Shock, we classified periods of hypotension with a given patient as any period of time where hypotoclic blood pressure was below 90. Using the center time of each of these periods of hypotension, we summed up all the fluids given to the patient between this time and an hour before. If the amount of fluids in this timespan exceeded 600ml, the patient was considered to have hypotension despite fluid intake. As such, if the patient was also determined to have Sepsis, we labeled the patient as positive for Septic Shock. The start time of that patient’s period of hypotension was used as the onset time.

Predicting Septic Shock

To do our predictive analysis, we used four existing mathematical models: a Logistic Regression model, a Decision Tree model, an Artificial Neural Network, and an existing mathematical models: a Logistic Regression model. To do our predictive analysis, we used four existing mathematical models: a Logistic Regression model, a Decision Tree model, an Artificial Neural Network, and an existing mathematical models: a Logistic Regression model.

Results

Each model yielded comparable results for the two data sets. To measure the performance of our models, we created for each a Receiver Operating Characteristic (ROC) curve.

Logistic Regression

GMC AUC: .75, MIMIC AUC: .8408

Decision Tree (RPART)

GMC AUC: .875, MIMIC AUC: .7451

GMC AUC: .8325, MIMIC AUC: .7385

Conclusion

We replicated models in the literature on the same publicly available data set on which they were developed. After getting these baseline results, which were similar to those in the literature, we applied the same models (with the same features and definitions) to privately held data. For a fixed model, its performance on the publicly available data and the privately held data were comparable. The quality of performance was especially good considering that the models were developed on demographic features of the individuals and only seven additional features. In some cases, our models were trained on larger data sets and also outperformed models in the literature.

The greatest challenge in the development of these models is the determination of which individuals have the right kind and right amount of data to be included in the models. In particular, identifying which patients had sepsis and when they had it (and septic shock if applicable) was essential for our experiments. The fact that in some cases we improved on existing results we think can be attributed to the choice that we made to identify patients as having sepsis in a way that is not based on the ICD-9 the patients have had.

For our next step, we will be examining how we might examine the paths of how patients move through the hospital, and use predictive techniques like Markov Chains on the GMC data.

References