Machine Learning Engineer Nanodegree

Capstone Proposal

Wenbo Wang

January 31st, 2018

Domain Background

Population aging, especially in developed world, is historically unprecedented, pervasive, and has profound impact on economy as well as many social aspects like healthcare and politics[1]. At the same time, healthcare spending in the United States already claims 17.1% of its gross domestic product (GDP), far exceeding other high-incoming countries[2]. To make sure that all nations are prepared for providing efficient and cost-effective healthcare in the foreseeable future, we have to come up with strategies both on the technological fronts and at different policy levels. We can innovate medical devices, design better pharmaceutical drugs, and improve hospital capacities, which directly add benefits to patient care. On the other hand, we also have to investigate critically whether current clinical practices/policies could be improved to maximize health and economic gains without further investing. For instance, expanding medical intensive care unit (ICU) capacity was not found to be effective in improve clinical outcomes for critically ill elderly patients[3]. In recent years, the national push for electronic medical record (EMR) has been the main driver behind rapid aggregation of gigantic volumes of detailed digitized health data[4]. It is no doubt that the wealth of health data generated by modern instruments and medical practitioners presented unprecedented opportunities as well as challenges for both doctors and scientists. With the highly heterogeneous data drew from targeted patient population, we can afford to take a closer look at the intertwined relations between illnesses and treatments. Ultimately, exploring EMR will allow us to customize treatment for optimal care of individual patients, making sure that each unique hospital admission receive the right and timely medical interventions at the right facility.

Problem Statement

One of the important health outcomes for ICU admissions is patient mortality. Accurate prediction of patient mortality scores means both better chance of survival for the patients and more efficient use of constrained ICU resources. Presently, patient mortality is assessed by evaluating severity scores such as Acute Physiology and Chronic Health Evaluation (APACHE) II scores, Simplified Acute Physiology Score (SAPS) II scores, etc[5,6]. These scores are usually based on a limited set of hand-picked predictors/variables and then used to calculate mortality probabilities[7]. In this capstone project, I am interested to find out whether use of machine learning methods could improve patient mortality prediction based on patients' health record in a de-identified EMR database. Here, a list of patient EMR information, which may include demographic data, lab analysis results, biometric records, and ICU data are selected as the preditors. The outcome of interest is patient mortality (0 survival and 1 death). Several classification methods are evaluated for their diagnostic ability by calculating the receiver operator curves (ROCs) and area under curve (AUC)values. The data source

is a publicly available database called "Medical Information Mart for Intensive Care", namely, the MIMIC-III set. Query commands to pull the dataset for mortality prediction and machine learning scripts are to be published on github under my account at the end of the project.

Datasets and Inputs

To answer the question of accurately predicting patient mortality based on health record, I used the MIMIC-III database. The MIMIC-III database was created and maintained by the Laboratory for Computational Physiology at Massachusetts Institute of Technology (MIT)and the department of Medicine at the Beth Israel Deaconess Medical Center (BIDMC)[8]. In brief, the original dataset contains 53,423 unique hospital admissions of adult patients admitted to ICU between 2001 and 2012. There are 38,597 adult patients with a median age of 65.8 years, 55.9% male, and in-hospital mortality of 11.5%. There are on average 4579 charted observation and 380 laboratory measurements for each hospital admission. The bioinformatics contained in MIMIC-III dataset is appropriate for this project because it has been proven successful in predicting mortality for certain subgroup patients with diagnostic utility surpassing that of severity scoring systems[9,10].

To assess MIMIC-III database, I took training course on respectful usage of clinical data and was subsequently granted access. Then I built a local database service for MIMIC-III under PostgreSQL v.10.1. I have also utilized author's github SQL scripts to build severity score tables for SAPS[11]. For this project, I plan to extract predictor variables based on previously published literature findings[10]. Most of the variables are laboratory analysis results, e.g. potassium levels, sodium levels, ICU data, e.g., total urine output during 24 hours, and demographic information such as age, sex, etc. The predicted outcome is the mortality, which is available in the admissions table with survival to hospital discharge coded as 0 and 1 otherwise. Only adult patients are included as the target group. There are around a total of ~24,000 adult patients and I only considered their first admission in this project. For the selected features, a subset of ~20 features will be used. For the selected dataset, about 12% of patients did not survive till hospital discharge. Because the MIMIC-III data is suited for nonrandomized retrospective studies, it is quite difficult to balance the training/test. Fortunately, the dataset is big and the data can be split randomly and check if the summary statistics for selected variables statistically different from each group. Class balance is maintain to make sure calibrated model can generalize to the targeted population. F1 score and precision/recall metrics instead of accuracy is calculated to check that algorithms are learning from imbalanced data.

Solution Statement

The commonly used SAPS scoring system adopted logistic regression for calculating probability of in-hospital mortality. The performance was useful but somehow mediocre with an AUC between 0.6 and 0.7. In this project, I intend to test a support vector machine (SVM) and a random forest (RF) classifier for predicting patient mortality. The SVM classifier adopts nonlinear kernel functions and is more tolerant to outliers, giving it good generalization performance. The RF classifier is very quick to train and requires little data preparation. Both algorithms

are suitable for analyzing MIMIC data, which is highly heterogeneous and may contain lots of missing or invalid inputs. To compare performance, filtered patient dataset is split into training and test test with a 3:1 ratio. Both ROC and AUC are calculated and compared. Other metrics such as precision and recall are also to be generated.

Benchmark Model

The SAPS, which is a standard illness severity scoring system is used as the benchmark model for fair comparison. The predicted mortality based on SAPS scores is calculated as:

$$Predicted\ Death\ Rate = \frac{e^{(Logit)}}{1+e^{(Logit)}} \quad (1)$$

Where Logit=7.7631+0.0737*SAPS+0.9971*ln(SAPS+1)[5].

Based on the SAPS derived predicted mortality rate, both prediction accuracy and AUC values can be calculated. Thus, those values could be used against their counterparts generated using SVM and RF claffiers by applying them to the same test dataset.

Evaluation Metrics

For this binary classification problem, I am going to use the following performance evaluation metrics in assessing training classifiers: Precision defines how much prediction mortality are actual death when compared to hospital records:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall defines the portion of actual death that has been successfully predicted:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

The F1 score is a weighted average of precision and recall, which is defined as :

$$F1 = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

The ROC curve and the AUC are calculated and compared agains that of the standard severity scoring system[12].

Project Design

In figure 1, I provided an overview about the entire workflow of data extraction, machine learning, and result presentation to be deployed in this project. All data entries in the MIMIC-III database that satisfy the filter conditions, e.g., adults (>18), minimum length of stay >72 hours, and complete health record during stay, are extracted using SQL script to pull the data. Once raw data is extracted. Additional data cleansing is performed by observing the histograms, calculating summary statistics, and performing data transforms if necessary. Aggregated metrics SAPS is computed using published sql code and extracted to calculate predicted mortality benchmarks. Summary statistics are calculated for all selected variables and used to decide if further transform is needed. For example, one-hot-code may be needed to transform some categorical variables.

The cleaned bioinformatics data is then split into a training set and a validation set with 25% of the data randomly selected and held out for independent test of model performance. The same training set is used for train both SVM and RF classifiers with 10-fold-cross-validation for early stopping and model selection. The hyperparameters for SVM and RF are subsequently tuned in the model refinement stage where a grid search or random search is employed to optimize model performance.

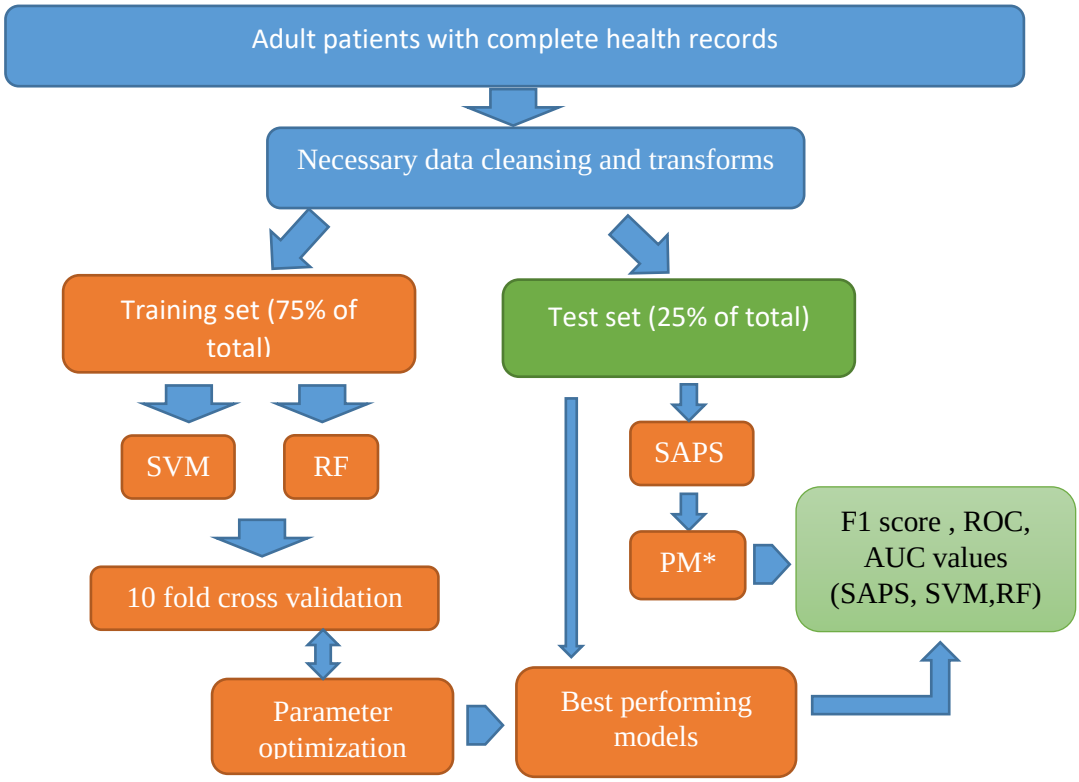

Figure 1. workflow for predicting patient mortality using MIMIC-III data

For SVM classifier, the parameters to be tuned include C, the penalty term, and gamma, the kernel coefficient. Random forest classifier have a lot more parameters to tune, which include the number of estimators, max depth , min samples per leaf, etc. Once parameters for an optimized model is ready, the best performing models are used to predict the independent test set and subsequently generate all performance statistics such as F1 score, AUC values, etc.
The goal is to find out whether a machine learning model could offer better prediction of patient mortality compared to the standard scoring system. Thus, ROC are generated along with AUC values for direct comparison among different models. The feature importance, which is available in RF classifier results can potentially offer some insights into the important biometric signal for clinicians to take actions in the ICU that maximize chances for patient survival.

References

1. Nations U. World population ageing: 1950-2050. *World Popul Ageing*.

2002;26(26). http://www.un.org/esa/population/publications/worldageing19502050/%5Cnpapers2://publication/uuid/B62F39F7-A14C-44CE-AEB2-96D9E003F8BE.
2. Squires D, Anderson C. U.S. Health Care from a Global Perspective. *Commonw Fund*. 2015. http://www.commonwealthfund.org/publications/issue-briefs/2015/oct/us-health-care-from-a-global-perspective.
3. Fuchs L, Novack V, McLennan S, et al. Trends in severity of illness on ICU admission and mortality among the elderly. *PLoS One*. 2014;9(4). doi:10.1371/journal.pone.0093234.
4. Che Z, Purushotham S, Khemani R, Liu Y. Interpretable Deep Models for ICU Outcome Prediction. *AMIA . Annu Symp proceedings AMIA Symp*. 2016;2016:371-380. http://www.ncbi.nlm.nih.gov/pubmed/28269832%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5333206.
5. Gall JR, Lemeshow S, Saulnier F. A New Simplified Acute Physiology Score (SAPS II) Based on a European/North American Multicenter Study. *JAMA J Am Med Assoc*. 1993;270(24):2957-2963. doi:10.1001/jama.1993.03510240069035.
6. Rogers J, Fuller HD. Use of daily Acute Physiology and Chronic Health Evaluation (APACHE) II scores to predict individual patient survival rate. *Crit Care Med*. 1994;22(9):1402-1405. doi:10.1097/00003246-199409000-00008.
7. Keegan MT, Gajic O, Afessa B. Severity of illness scoring systems in the intensive care unit. *Crit Care Med*. 2011;39(1):163-169. doi:10.1097/CCM.0b013e3181f96f81.
8. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3. doi:10.1038/sdata.2016.35.
9. Danziger J, Chen KP, Lee J, et al. Obesity, Acute Kidney Injury, and Mortality in Critical Illness. *Crit Care Med*. 2016;44(2):328-334. doi:10.1097/CCM.0000000000001398.
10. Celi LA, Galvin S, Davidzon G, Lee J, Scott D, Mark R. A Database-driven Decision Support System: Customized Mortality Prediction. *J Pers Med*. 2012;2(4):138-148. doi:10.3390/jpm2040138.
11. Johnson AE, Stone DJ, Celi LA, Pollard TJ. The MIMIC Code Repository: enabling reproducibility in critical care research. *J Am Med Informatics Assoc*. 2017. doi:10.1093/jamia/ocx084.
12. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2012;12:2825-2830. doi:10.1007/s13398-014-0173-7.2.