Mortality and Death Time Prediction Models using MIMIC-III

Suiki LAU, Online Master of Science in Computer Science
Georgia Institute of Technology, Atlanta, Georgia, USA

Abstract

Most ICU mortality prediction models in the literature were designed for at least 24 hours or 48 hours after ICU admission to provide real-time or retrospective prediction on patients' mortality. This study proposes a machine learning approach to address the task of predicting in-hospital mortality in the early stage of ICU stay using 6-hour timeframe. If a patient is predicted dead, the model would further provide an estimate of death hours since ICU admission. The aim of the study is to identify high-risk patients who might be dead within hours or days since ICU admission. Data were extracted for the first 6 hours, 12 hours or 24 hours since ICU admission for each ICU stay from MIMIC-III database. The extracted features includes 45 physiological variables such as heart rate, blood pressure, Glasgow coma scale, and demographic features such as gender, age, ethnicity. Although there are many missing values in the first 6-hour of ICU data, we have demonstrated in the study a feasible and novel framework to predict in-hospital mortality and death time. The mortality prediction model trained on the first 6-hour ICU data has competitive performance (AUROC 0.88), whereas the death time multiclass classifier offers an effective base (micro-average AUROC 0.76) to provide a rough estimate of death hours since ICU admission.

Introduction

An intensive care unit (ICU) provides intensive treatment medicine for patients with severe and life-threatening illness and injuries, such as trauma, multiple organ failure, sepsis or directly transferred from emergency department. ICUs have higher staff-to-patient ratio then normal wards in order to provide intensive care and comprehensive monitoring to severe patients. Hence, ICUs generate a massive amount of electronic healthcare records which are useful to predict patients' disease status and the amount of healthcare needed. These records should provide many explanatory variables including demographic, physiological, vital signs and laboratory test variables in predicting patients' in-hospital mortality. A machine learning approach can automate the process in extracting useful features from the massive information in the database, and help improve the efficiency and quality of the healthcare system.

The Medical Information Mart for Intensive Care III (MIMIC-III)[1–3] is a freely-accessible database comprising de-identified electronic healthcare records of over 60,000 ICU stays for around 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The database consists of rich information about patients' demographic characteristics, such as gender, age, ethnicity, admission type, and various in-hospital measurements, lab tests, procedures and medication of ICU patients over the time. The database provides data from two electronic healthcare record systems, namely the CareVue (from 2001 to 2008) and MetaVision (from 2008 to 2012), which collect and store data differently.

Over the past few decades, several ICU scoring systems[4,5] have been developed for mortality prediction using rule-based method or data mining approach. Some standard scoring systems include Acute Physiology And Chronic Health Evaluation (APACHE), Simplified Acute Physiology Score (SAPS) and Sepsis-related organ Failure Assessment Score (SOFA). APACHE[6–8] is a severity scoring systems designed to provide a morbidity score for a patient. A predicted mortality can be derived from this score. SAPS[9] was designed to predict morbidity for a particular patient by comparing the outcome with other patients or a group of patients by comparing the outcome with another group of patients. SOFA[10] provides a daily score to track a person's status during an ICU stay to determine the extent of a person's organ function or rate of failure.

With the advancements in parallel and distributed computing and machine learning, better mortality prediction models trained on larger input features have been developed using machine learning approach. Desautels et al.[11] introduced InSight, a machine learning classification system that uses multivariable combinations of easily obtained patient data (vitals, peripheral capillary oxygen saturation, Glasgow Coma Score, and age), to predict sepsis using MIMIC-III. Ghassemi et al.[12] applied Latent Dirichlet Allocation to free-text hospital notes to make mortality prediction using MIMIC-II. Hrayr et al.[13] applied Recurrent Neural Network and provided four clinical prediction benchmarks using

MIMIC-III. Pirracchio et al.[14] applied ensemble methods to improve mortality prediction in the ICU. Awad et al.[15] provided early hospital mortality prediction of ICU patients, i.e. first 6 hours since ICU admissions, using an ensemble approach.

Objectives

Most models in the literature were designed for at least 24 hours or 48 hours after ICU admission to provide real-time or retrospective prediction on patients' mortality. In this study, we propose to a two-phase model framework to address the task of predicting the mortality and also death hours in the early stage of ICU stay using 6-hour timeframe. If a patient is predicted dead in the first phase, the model would further provide an estimate of death hours since ICU admission in the second phase. The aim is to identify high-risk patients who might be dead within hours or days since ICU admission. Data were extracted for the first 6 hours, 12 hours or 24 hours since ICU admission for each ICU stay from MIMIC-III database. Multiple models were trained on the extracted features of the study population for the specified timeframe. The model results were then compared and discussed in the study.

Methods

Exploratory Data Analysis

The original dataset in MIMIC-III database consists of 61,532 distinct ICU stays of 46,520 unique patients. To form our study population, we have excluded ICU stays less than one hour to remove fuzziness in data due to unusual short stays and only consider adult patients with age between 16 and 89. The final study population covered 49,632 ICU stays of 36,343 patients. Multiple machine learning models in this study have been trained and evaluated based on this study population. Table 1 provides summary statistics of the study population.

Variables	Statistics
Age	Mean 62.61
Gender	Male 57.79%
Ethnicity	White 71.00%
Admission type	Emergency 82.31%
Number of ICU stays	Mean 1.37
In-hospital mortality ratio	11.62%

Table 1: Summary statistics of the study population

Among 49,633 ICU stays, there are 5,766 in-hospital mortality. After filtering out the ICU stays with negative death time since ICU admission (which is likely an administrative error resulting in an incorrect ICU admission or incorrect death time), 5,718 in-hospital mortality were resulted. The average death time since ICU admission is 9.57 days, maximum death time is 206.38 days and minimum death time is 0 day. Figure 1 shows the distribution of death time in hours since ICU admission for in-hospital mortality.

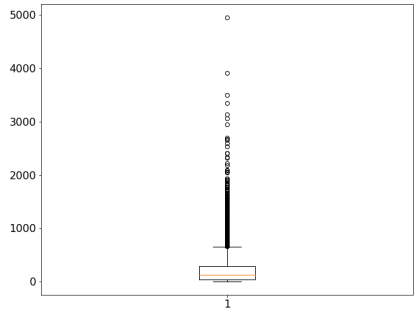

Figure 1: Box plot of death time in hours since ICU admission

Feature Extraction

Our study aims to predict mortality and death time in the early stage of ICU stay. For each ICU stay, we have extracted data from the first 6 hours, 12 hours and 24 hours since ICU admission. We first extracted patients' features in every hour of their ICU stay until they were discharged from the ICU. We then aggragated the feature values in the specified timeframe. There are altogether 123 extracted features covering 5 static variables and 40 physiological variables. Reference has been made to this code repository[16] when we constructed the features.

The static variables includes admission type, total number of previous and current ICU stays, and demographic features such as age, gender and ethnicity. The temporal data of physiological variables includes patients' vital signs such as heart rate and blood pressure, Glasgow coma scale, blood gases and chemistry values, laboratory results and urine output. Most of the temporal variables were aggregated by maximum, minimum, and average during the specified timeframe (6 hours, 12 hours or 24 hours), except that urine output was aggregated by sum. Table 2 provides the list of extracted features used in the study.

Categories	**Variables**	**Extracted features**
Demographic and static features	Age, Gender, Ethnicity, Admission type, Number of ICU stays	N/A
Vital signs	Heart rate, Systolic blood pressure, Diastolic blood pressure, Mean blood pressure, Respiratory rate, Temperature, Peripheral capillary oxygen saturation, Glucose	Minimum, Maximum, Mean
Glasgow coma scale	Glasgow coma scale (GCS), GCS components (motor, verbal, eyes)	Minimum, Maximum, Mean
Blood gases and chemistry values	Base excess, Carboxyhemoglobin, Methemoglobin, Partial pressure of oxygen, Partial pressure of carbon dioxide, pH, Ratio of partial pressure of oxygen to fraction of oxygen inspired, Total carbon dioxide concentration	Minimum, Maximum, Mean
Lab results	Anion gap, Albumin, Immature band forms, Bicarbonate, Bilirubin, Calcium, Creatinine, Chloride, Hematocrit, Hemoglobin, Lactate, Platelet, Potassium, Partial thromboplastin time, International Normalized Ratio, Sodium, Blood urea nitrogen, White blood cell count	Minimum, Maximum, Mean
Urine output	Urine output	Sum

Table 2: List of extracted features

Model Architecture

We propose a two-phase model framework to predict in-hospital mortality and death time in hours. In Phase 1, a binary classifier was trained using the 123 extracted features listed in Table 2 to predict in-hospital mortality. In Phase 2, a multiclass classifier was trained on the same set of extracted features to predict death time in hours since ICU admission for the predicted dead patients in Phase 1.

In Phase 1, 49,632 ICU stays of our study population (filtered by age between 16 and 89, and length of ICU stay more than 1 hour) were split into 80% training set and 20% test set. Random forest classifiers were trained on the training set to predict the in-hospital mortality label (around 12% were dead) using 6-hour, 12-hour and 24-hour data respectively. Hyperparameter tuning was performed using grid search on 5-fold cross-validation (CV) of the training set. The model performance of the best classifier resulted from the grid search under 6-hour, 12-hour and 24-hour scenarios were then compared and evaluated on the test set.

In Phase 2, we filtered out dead patients with negative death time. A total of 5,718 ICU stays of dead patients were split into 80% training set and 20% test set. We then label each data to one of the three specified classes. Table 3 shows the distribution of data across the three classes and their defintions. Random forest multiclass classifiers were then trained on the training set to predict the death time label using 6-hour, 12-hour and 24-hour data respectively. Hyperparameter tuning was performed using grid search on 5-fold cross-validation (CV) of the training set. The model performance of the best classifier resulted from the grid search under 6-hour, 12-hour and 24-hour scenarios were then compared and evaluated on the test set.

Class	Description	Number of data
Class 0	death time hours $<$ 24	838
Class 1	24 $\leq$ death time hours $<$ 24*7	2517
Class 2	24*7 $\leq$ death time hours	2363

Table 3: Distribution of data across the three classes

Evaluation

The dataset was split into 80% training set and 20% test set. Hyperparameter tuning was done on 5-fold CV of the training set and the final evaluation of model performance was done on the test set. Multiple machine learning models in the study were evaluated and compared using the Area under the Receiver Operating Characteristic curve (AUROC) on the test set. The ROC curve is the true positive rate against the false positive rate at various threshold settings. AUROC provides a single measure of the diagnostic ability of a binary classifier as its discrimination threshold is varied. We have used AUROC to compare the model performance of the binary classifiers in Phase 1.

As for multiclass classifier, we have used micro-average and macro-average AUROC to evaluate the model performance in Phase 2. In gist, macro-average treats all classes equally, computes the metric independently for each class and then take the average, whereas micro-average aggregates the contributions of all classes to compute the average metric. In a multi-class classification setup, micro-average is preferable when there is class imbalance. Apart from the primary metric of AUROC, accuracy, confusion matrix, precision, recall and F1-score were also used during the model testing stage to provide a full picture of model performance.

Implementation

The study consists two stages of implementation:

- Feature engineering using Apache Hive 2.1.0 on Microsoft Azure remote cluster (2 head nodes and 1 worker node, each with 200 GB space, 14 GB RAM, and 4 processors)
- Machine learning using Python 3.6 on a local cluster (500 GB space, 16 GB RAM, 4 GB GPU and 4 processors)

The decompressed dataset of MIMIC-III requires around 50 GB of space. We chose big data tool like Apache Hive to perform data preprocessing and feature engineering since the dataset is quite large. Data output in the first stage

is then used as feature input for the model training in the second stage. We also used Python and packages such as Pandas and Scikit-learn for efficient model testing, hyperparmeter tuning and model evaluation.

A set of reproducible code (including Hive and SQL scripts, Python notebooks), detailed instruction about implementation (README file), and presentation slides have been provided along with the paper. The code and the presentation video can also be found in the following webpages.

Experimental Results

Table 4 compares the model performance of the random forest classifiers separately trained using 6-hour, 12-hour and 24-hour data in Phase 1. The results show that data aggregation over wider timeframe gives better result. Specifically, the random forest classifiers trained on 6-hour, 12-hour, 24-hour data have AUROC 0.88, 0.90, 0.92 on the test set respectively. Figure 2 compares the ROC curves of the three classifiers.

Models	AUROC on 5-fold CV	AUROC on test set
Random forest classifier trained on 6-hour ICU data	0.883	0.883
Random forest classifier trained on 12-hour ICU data	0.901	0.902
Random forest classifier trained on 24-hour ICU data	0.920	0.919

Table 4: Model performance of random forest classifiers trained on various timeframe in Phase 1

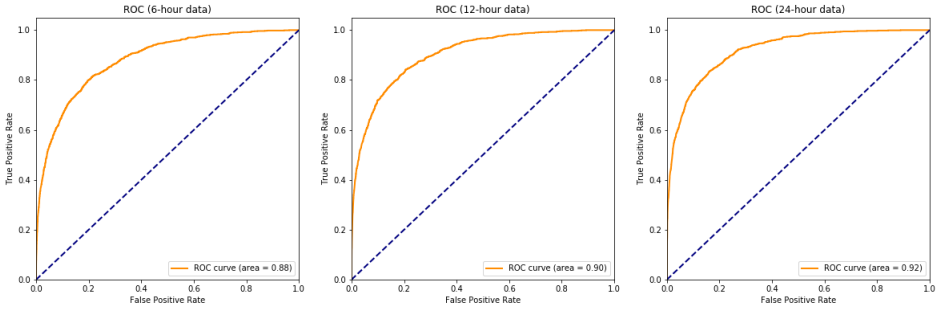

Figure 2: ROC curves of random forest classifiers evaluated on test set in Phase 1

Table 5 compares the model performance of random forest multiclass classifiers separately trained using 6-hour, 12-hour and 24-hour data in Phase 2. The results show that data aaggrgation over 24-hour timeframe gives slightly better result in terms of macro-average AUROC than those of 6-hour and 12-hour. Specifically, random forest classifier trained on 6-hour, 12-hour, 24-hour ICU data have the same micro-average AUROC 0.76 and macro-average 0.74, 0.73, 0.78 on the test set respectively. Figure 3, figure 4, figure 5 compare the micro-averge, macro-average and ROC curves of individual classes for the three multiclass classifiers respectively.

Models	Micro-average AUROC on test set	Macro-average AUROC on test set
Random forest multiclass classifier trained on 6-hour ICU data	0.76	0.74
Random forest multiclass classifier trained on 12-hour ICU data	0.76	0.73
Random forest multiclass classifier trained on 24-hour ICU data	0.76	0.78

Table 5: Model performance of random forest multiclass classifiers trained on various timeframe in Phase 2

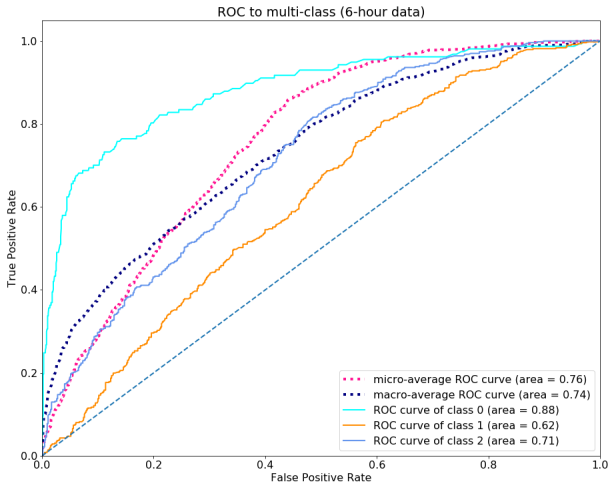

Figure 3: ROC curves of random forest classifier (6-hour data) evaluated on test set in Phase 2

Discussion

MIMIC-III is a rich source of electronic healthcare records comprising a diverse range of static data and high temporal resolution data for a large population of ICU patients. In many standard scoring systems in ICU setting, manual feature engineering is difficult and requires expert knowledge. The need of manual feature engineering can be obviated by automated feature learning using machine learning approach. We have shown that by combining static features such as patient's' demographic information and dynamic features such as physiological variables measured in ICU, we could train a effective model to predict in-hospital mortality in the early stage of ICU stay. The predictive performance of the model trained on the first 6-hour ICU data has competitive performance (AUROC 0.88) with the same framework trained on the first 12-hour or 24-hour ICU data (AUROC 0.90 and 0.92 respectively). We then showed that the death time multiclass classifier trained on the first 6-hour ICU data offers an effective base (micro-average AUROC 0.76) to provide a rough estimate of death hours since ICU admission. The result is competitive to those models using the first 12-hour or 24-hour ICU data (micro-average AUROC 0.76).

The framework demonstrated in the study provides a base for potential improvement. One possible way to improve the model performance is to identify more static and dynamic features and combine them in an effective way in the feature engineering stage. Specifically, we have aggregated dynamic features over a specific timeframe, like the first 6-hour since ICU admission, in our study. This indeed might give up quite a lot of temporal information. Methods such as Long Short Term Memory (LSTM), a type of Recurrent Neural Network, which is designed to handle sequence dependencies for time-series prediction problem, could be applied to handle temporal data. In our proof-of-concept stage, we have fitted LSTMs to some dynamic features, such as heart rate and blood pressure, and ensembled them

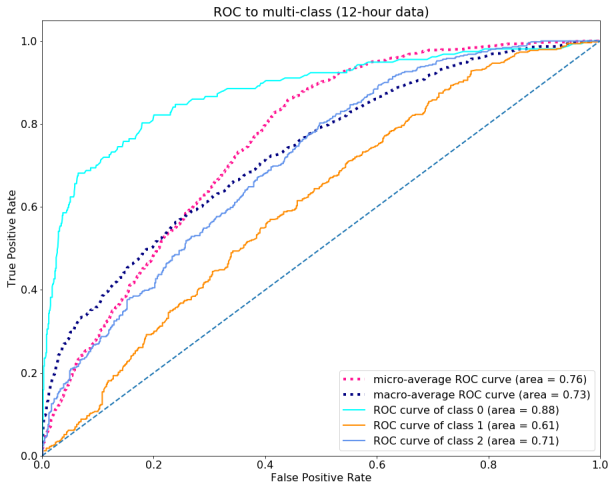

Figure 4: ROC curves of random forest classifier (12-hour data) evaluated on test set in Phase 2

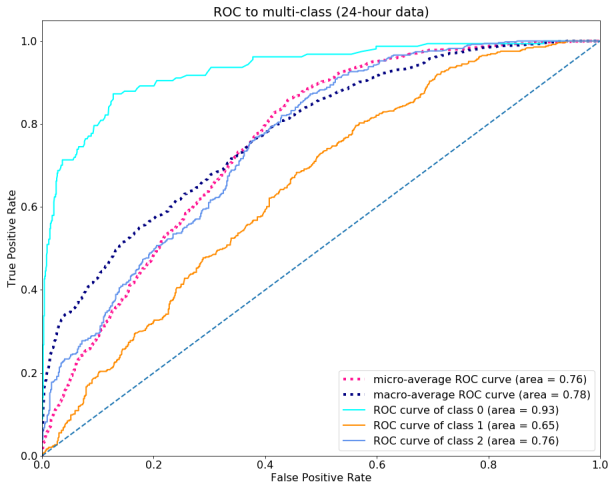

Figure 5: ROC curves of random forest classifier (24-hour data) evaluated on test set in Phase 2

with random forest classifiers fitted on other static or less temporal data. However, the model performance is not satisfactory due to many missing values in the time series of the dynamic features. Finding an effective algorithm to interpolate missing time-series data becomes crucial to constructing a robust predictive model.

Conclusion

The results show that although the models trained on the data in the first 24-hour since ICU admission give better performance, the first 6 hours of ICU data provides enough information for mortality prediction and a rough estimate of death hours since ICU admission. The proposed framework provides a base to identify high-risk patients who might be dead within hours or days since ICU admission in the early stage of ICU stay, and there are potential avenues for improvement.

References

1. A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. MIMIC-III, a freely accessible critical care database. Sci Data, 3:160035, 24 May 2016.
2. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/content/101/23/e215.full]; 2000 (June 13).
3. Johnson, Alistair EW, David J. Stone, Leo A. Celi, and Tom J. Pollard. The MIMIC Code Repository: enabling reproducibility in critical care research. Journal of the American Medical Informatics Association (2017): ocx084.
4. Soares M, Fontes F, Dantas J, Gadelha D, Cariello P, Nardes F, et al. (2004). Performance of six severity-of-illness scores in cancer patients requiring admission to the intensive care unit: a prospective observational study. Crit Care. 8 (4): R194203. doi:10.1186/cc2870. PMC 522839Freely accessible.
5. Strand K, Flaatten H (2008). Severity scoring in the ICU: a review. Acta Anaesthesiol Scand. 52 (4): 46778. doi:10.1111/j.1399-6576.2008.01586.x.
6. Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985). APACHE II: a severity of disease classification system. Critical Care Medicine. 13 (10): 81829. doi:10.1097/00003246-198510000-00009. PMID 3928249. (This is the first published description of the APACHE II scoring system)
7. Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE (1981). APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. Critical Care Medicine. 9 (8): 5917. doi:10.1097/00003246-198108000-00008.
8. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, Sirio CA, Murphy DJ, Lotring T, Damiano A, et al. (1991). The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest. 100 (6): 161936. doi:10.1378/chest.100.6.1619.
9. Jean-Roger Le Gall, MD; Stanley Lemeshow, PhD; Fabienne Saulnier, MD. (1993). A New Simplified Acute Physiology Score (SAPS II) Based on a European/North American Multicenter Study. JAMA. 1993;270:2957-2963 This is the first published description of the scoring system
10. Vincent JL, Moreno R, Takala J, Willatts S, De Mendona A, Bruining H, Reinhart CK, Suter PM, Thijs LG. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. Intensive Care Med 1996 Jul;22(7):707-10.
11. T. Desautels, J. Calvert, J. Hoffman, M. Jay, Y. Kerem, L. Shieh, D. Shimabukuro, U. Chettipally, M. D. Feldman, C. Barton, D. J. Wales, and R. Das. Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach. JMIR Med Inform, 4(3):e28, 30 Sept. 2016.
12. M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits. Unfolding Physiological State: Mortality Modelling in Intensive Care Units. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 14, pages 7584, New York, NY, USA, 2014. ACM.
13. H.Harutyunyan, H.Khachatrian, D.C.Kale,andA.Galstyan. Multitask learning and benchmarking with clinical time series data. arXiv preprint arXiv:1703.07771, 22 Mar. 2017.
14. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der LAAN MJ. Mortality prediction in the ICU: can we do better? Results from the Super ICU Learner Algorithm (SICULA) project, a population-based study. The Lancet Respiratory medicine. 2015;3(1):42-52. doi:10.1016/S2213-2600(14)70239-5.
15. Awad A1, Bader-El-Den M2, McNicholas J3, Briggs J1. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. Int J Med Inform. 2017 Dec;108:185-195. doi: 10.1016/j.ijmedinf.2017.10.002. Epub 2017 Oct 5.
16. Github Repository available at https://github.com/alistairewj/mortality-prediction/tree/master/queries