CinC Challenge: Predicting In-hospital Mortality of Intensive Care Unit by Analyzing Histogram of Medical Variables under Cascaded Adaboost Model

Chucai Yi[1,2], Yi Sun[1], and Yingli Tian[1,2*]

[1]Dept. of Electrical Engineering, The City College of New York, City Univ. of New York, USA
[2]Dept. of Computer Science, The Graduate Centre, City Univ. of New York, USA

Abstract

*In this paper, we develop an effective framework to predict in-hospital mortality (IHM) during intensive care unit (ICU) stay, on the basis of specific medical variables. This work involves both binary mortality predictions and mortality risk estimates, corresponding to Event-1 and Event-2 of the Computing in Cardiology (CinC) Challenge 2012. Our proposed framework contains 1) feature extraction from medical variables by linear interpolation, histogram analysis, and temporal analysis; and 2) mortality classifier learning under Cascaded Adaboost learning model. A released dataset **set-a** of ICU medical records is used as training set, where cross validation is performed to evaluate our proposed framework. Our framework achieves Event-1 Score1 0.806 and Event-2 Score2 24.00, which outperform those obtained from SAPS-1 score (Score1 0.296 and Score2 68.39) on the same dataset. Over another dataset **set-b**, our framework obtains Event-1 Score1 0.379 and Event-2 Score2 5331.15.*

1. Introduction

Medical variable measurements from Intensive Care Unit (ICU) play an important role in clinical research.The measurement values of medical variables during ICU stay reflect patients' physical conditions and variations which are able to predict disease progression.

In CinC-Challenge 2012 [3], disease progression is defined as in-hospital mortality rate, that is, observing whether the patient dies during ICU stay or not. However, it is a challenging task to model an accurate relationship between mortality rate and the observed medical variables due to the following reasons: 1) disease progression is analyzed based on a group of medical variables at specific time; 2) the training set does not provide a uniform measure plan for all patients, that is, different patients generate different medical measurements at the same time; 3) it is impossible to measure all medical variables with high frequency during all ICU stay, and the missing data will bring more difficulties in the statistics of medical variables.

In this paper, we propose an algorithm to estimate in-hospital mortality rate according to the measurement values of medical variables during a 48-hour ICU stay. A mortality classifier is generated from training set.

2. Data

The medical data released for CinC-Challenge-2012 [3] consists of three datasets: *set-a*, *set-b*, and *set-c*. Each of them contains medical records of 4000 patients during ICU stay. *Set-a* is used as training set because its outcomes of in-hospital mortality are provided as ground truth. *Set-b* and *set-c* without ground truth labels are used as testing sets. The medical records in all the three datasets cover the same group of 37 medical variables, which are measured in the first 48 hours of ICU stay. In the following sections, we define a sequence of records from one patient during 48-hour ICU stay as an ICU sample.

In an ICU sample, 37 medical variables are generally measured within the 48 hours. In addition, each ICU sample has 6 associated descriptors, *RecordID*, *Age*, *Gender*, *Height*, *ICUType*, and *Weight*. Thus an ICU sample consists of a total of 43 variables. Some variables have 2 or more measurement values within the 48-hour ICU stay, while some are not measured at all for a patient. It results in different data dimensions among different ICU samples.

In training set *set-a*, each ICU sample is assigned a ground truth label. It provides the total length of ICU stay and the survival length of each patient. If the survival length is less than the ICU stay length, the patient is denoted as in-hospital death. In the rest of this paper, an ICU sample is defined as a positive ICU sample if the patient dies in hospital, and otherwise it is defined as a negative ICU sample as survival. The *set-a* contains 554 positive samples and 3446 negative samples.

3. Method

3.1. Data Interpolation

To normalize all ICU samples into feature vectors with a fixed dimensionality, we employ linear interpolation to complement missing measurement values of the medical variables. 37 medical variables and 2880 minutes (48-hour ICU stay) are defined as interpolation domain.

According to medical knowledge, we employ a normal value for each of the 37 original medical variables.

For each medical variable, a 1×2880 zero-element vector is initialized to represent each minute of the 2880-minute ICU stay. Then we fill in the measurement values according to their measurement time. Next, linear interpolation as Eq. (1) is performed between each pair of neighbouring nonzero elements.

$$v_i = \frac{d_i - d_L}{d_H - d_L}(v_H - v_L) + v_L \quad (1)$$

where d_L and d_H denote the indices of two neighbouring nonzero elements on the vector, and v_L and v_H are their corresponding values. v_i from interpolation is the value at the index d_i. Besides, all the zero elements in the ends of the vector are assigned the values of their nearest neighbouring nonzero element.

As mentioned above, some medical variables may not be measured during ICU stay. All the 2880 elements in its vector are zero without any information for interpolation. In this case, we assign the predefined normal value of the medical variable.

The linear interpolation is applied to all the 37 medical variables, and each of them obtains a 1×2880 vector. The 37 vectors contain all information of an ICU sample, which will be used as features to predict the in-hospital mortality.

3.2. Feature Normalization

To reduce computational complexity of feature extraction and classifier learning, we reduce the feature vector of each medical variable by calculating average measurement values.

For each medical variable, 2880 elements of the feature vector represent the corresponding measurement values of 2880 minutes during the 48 hours. We calculate a mean value for every 60 minutes, and cascade them into a feature vector in 48 dimensions of the medical variable with less dimensions. Each mean value denotes an average measurement of medical variable within an hour.

The reduction is applied to all the 37 medical variables, and we obtain a reduced feature vector in a total of $37 \times 48 = 1776$ dimensions. Over the 4000 samples in training set *set-a*, a 4000×1776 feature matrix is generated with each column corresponding to measurement values of a medical variable within an hour, and each row represent an ICU sample. Thus we define each row as a measurement-based feature vector. Feature vectors of all training samples from *set-a* can be merged into a measurement-based feature matrix.

3.3. Histogram Analysis

On the basis of the measurement-based feature matrix as described above, we model the distributions of medical variables by histogram analysis. It includes direct-histogram mapping and difference-histogram mapping.

3.3.1 Direct-Histogram Mapping

A column of the measurement-based feature matrix represents all 4000 measurement values of a medical variable in an hour. We generate a 72-bin histogram from each column of the feature matrix. This histogram represents characteristic distributions of the medical variable at this time, defined as direct-histogram, as shown in Figure 1.

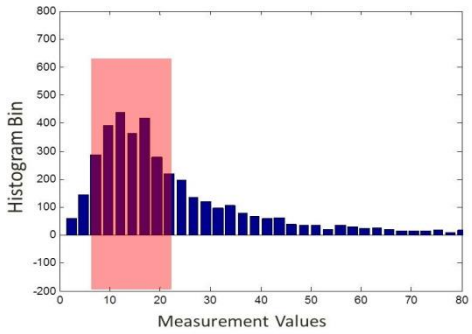

Figure 1. Direct-Histogram of blood urea nitrogen at the 20th hour, obtained from all training samples.

The elements of the measurement-based feature matrix are associated with the statistical values in the direct-histograms. For each element in the measurement-based feature matrix, we first check its column and the corresponding direct-histogram. Then the histogram bin of the element is calculated according to its value. Next, at this element, the original measurement value is replaced by the bin value of the direct-histogram. In this process, the measurement-based feature matrix is transformed into a statistic-based feature matrix. Similarly, each row of this statistic-based feature matrix represents an ICU sample. It contains statistical information of all medical variables over the ICU samples, so it gives more discriminatory power in the process of in-hospital mortality prediction.

3.3.2 Difference-Histogram Mapping

In addition to the direct statistics of medical variables, the measurement difference between the positive samples (death) and the negative samples (survival) can be adopted to analyze the in-hospital mortality.

Instead of histogram generation from all samples, two 72-bin histograms are calculated respectively from the positive samples and negative samples, as shown in Figure 2. We name them as positive histogram and negative histogram. Considering the difference between the numbers of positive samples and negative samples, we normalize the positive histogram by multiplying

$3446/554 = 6.22$ to each bin. To obtain statistics of the measurement difference, we calculate a difference-histogram by subtracting negative histogram from positive histogram, as shown in Figure 2. From the difference histogram, we infer that a greater bin value results in a higher likelihood of in-hospital death. The normalization of positive histogram enables positive values in difference histogram to represent high probability of in-hospital death.

Then we employ the same process as Section 3.3.1 to transform the measurement-based feature matrix into a statistic-based feature matrix according to difference-histogram. The two statistic-based feature matrices have the same size 4000×1776 as the original measurement-based feature matrix, and we fuse them vertically into a 4000×3552 statistic-based feature matrix. Each ICU sample has a row feature vector in 1×3552 dimensions.

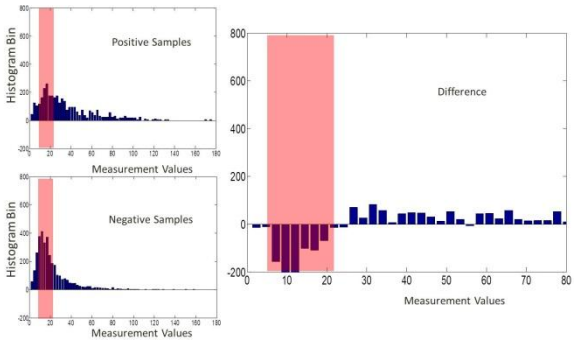

Figure 2. Left column shows the histogram of blood urea nitrogen at the 20th hour, obtained from positive samples and negative samples respectively. Right column presents the difference-histogram, obtained by the difference between the two left histograms.

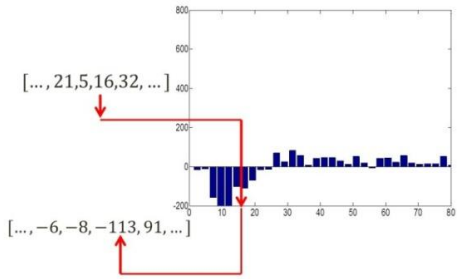

Figure 3. Mapping from measurements to statistics in histogram.

3.4. Temporal Analysis

Apart from the statistics of instant measurement values, the temporal variations of the medical variables also play an important role in morality prediction, as shown in Figure 4. It shows that temporal variation of a negative sample is smoother and closer to normal range.

We divide the ICU stay into two parts in equal length of 24 hours, and calculate the difference of their mean measurement values. As mentioned in Section 3.2, feature vector of each ICU sample is reduced into 48 dimensions, corresponding to the 48 hours during ICU stay. We divide this vector into N parts, and calculate the difference values at every $N/2$ step (see example in Figure 5 where $N = 4$). Then $N/2$ difference values are obtained as temporal features. In our experiments, we set $N = 8$, 4, and 2, and fuse all obtained difference values into a 7 dimensional vector. Combining the temporal features of all 37 medical variables, we obtain a feature vector in 259 dimensions. It demonstrates the temporal variations of an ICU sample.

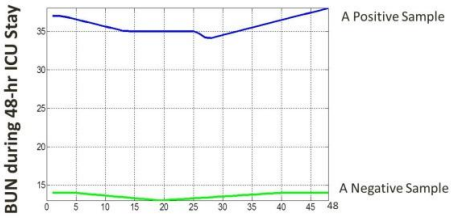

Figure 4. Temporal variation of blood urea nitrogen over positive sample (top curve) and negative sample (bottom curve).

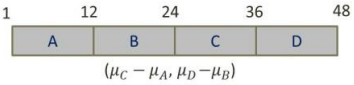

Figure 5. Feature vector of a sample is divided into 4 parts, and we calculate the part difference for every 2 steps, where μ represents the mean value of a part.

3.5. Cascaded Adaboost Model for Mortality Classification

For an ICU sample, we cascade its two feature vectors respectively obtained from histogram analysis and temporal analysis, and obtain a fused feature vector in $3552 + 37 = 3589$ dimensions. In addition to the 37 medical variables, the variable of "*Age*" also plays an important role in the prediction of in-hospital mortality. Thus we append the "*Age*" value as an additional medical measurement, and extend the feature vector into 3690 dimensions. This feature vector will serve as observation data of the ICU sample in the process of mortality classifier learning by a Cascaded Adaboost Model.

Cascaded Adaboost model was proved to be an effective machine learning algorithm in real-time face detection [4]. The training process is divided into several stages. Each stage is a decision-tree-based boost model. It performs an iterative selection of weak classifiers, based on the observations of all positive samples and the negative samples that are incorrectly classified in previous stages. The selected weak classifiers are integrated into a strong classifier by weighted combination [1]. The strong classifier will be used for predicting the risk of in-hospital death and survival.

The involved weak classifier h in the process of adaboost learning is defined by three parameters, which are column index j, threshold T, and polarity ρ. f_{*j}

denotes the *j*-th column of the feature matrix that contains all values of a medical variable at a time instant. At the column *j* of the feature matrix, we generate 48 thresholds that are uniformly distributed from minimum to maximum. Each threshold has 2 polarities as $\rho \in \{-1,1\}$. One polarity indicates that the *j*-th column value of positive sample is larger than the threshold while that of negative sample is smaller than the threshold, and the other is on the contrary, as presented in Eq. (2).

$$h(f_{i*}) = \begin{cases} pos & \rho f_{ij} \geq \rho T \\ neg & \rho f_{ij} < \rho T \end{cases} \quad (2)$$

where f_{ij} denotes the element at the *i*-th row and *j*-th column of the feature matrix, and f_{i*} denotes the *i*-th row vector that corresponds to the *i*-th ICU sample.

At the beginning of each stage, we prepare the training samples by selecting all positive samples and the hard negative samples, which are incorrectly classified by all the previous stages. The iteration processes stops when 99.5% of positive samples and 50% of negative samples are correctly classified. The Adaboost classifiers obtained from all stages are cascaded into the final mortality classifier. When the feature vector of a testing ICU sample is input into the final classifier, it is classified as in-hospital death if all cascaded strong classifiers determine it a positive sample, and otherwise it is in-hospital survival.

In the classification process, the weighted combinations of weak classifier output and the stage threshold are used to calculate the mortality risk for score2.

4. Results and Discussions

We perform experiments of mortality prediction on the ICU medical data. In Event-1, two measures sensitivity (*Se*) and positive predictivity (*P+*) are defined based on true positive (*TP*), false positive (*FP*), true negative (*TN*), and false negative (*FN*) [3]. The Score1 is defined as the minimum value of *Se* and *P+*. In Event-2, a mortality risk is estimated for each testing sample, and the Hosmer-Lemeshow *H* statistic algorithm is employed to calculate the Score2.

4.1. Evaluation over Training Set *set-a*

Set-a contains 4000 ICU samples, of which 554 are positive samples that are death during ICU stay and the other 3446 are negative samples that are survival in hospital. Each sample generates a feature vector from histogram analysis and temporal variation as described in Section 3.

We evenly divide the 4000 samples into three parts, and each part contains 185 positive samples and 1149 negative samples. Two of the three parts are employed to train the mortality classifier, and the resting one part is used for testing. We calculate their mean value as the cross validation result of score1 over ***set-a***. The result is *0.56*, which is higher than SAPS-1 [2] score *0.296*.

Next, we evaluate the mortality classifier obtained from all samples of *set-a* over itself. We can obtain the over-fitting results score1 *0.806* and score2 *24.00*.

4.2. Evaluation over Testing Set *set-b*

From all samples of training set ***set-a***, we learn a mortality classifier and then evaluate it over testing set ***set-b***. This testing set also contains 4000 samples, and we generate an observation feature vector for each sample by analyzing the direct-histogram and difference-histogram. According to the black box evaluation online, our best results over ***set-b*** are score1 *0.379* and score2 *5331.15*. The score2 evaluation based on Adaboost output is not well generalized to testing set.

5. Conclusion

In this paper, we have designed an effective algorithm to predict the in-hospital mortality and estimated the mortality risk of the patients during ICU stay. Measurement statistics and temporal variations of the medical variables are adopted to extract features from the training set. Each patient in the training set is considered as an ICU sample. We generate observation data of the sample as a feature vector, obtained by mapping the medical measurement values into the bin values of direct-histogram and difference-histogram. Then the observation feature vectors of all samples are input into the Cascaded Adaboost model to learn a mortality classifier. Our framework achieves Event-1 Score1 0.806 and Event-2 Score2 24.00, which outperform those obtained from SAPS-1 score [2] (Score1 0.296 and Score2 68.39) on the same dataset.

References

[1] Y. Freund and R. Schapire. Experiments with a new boosting algorithm. In *Int. Conf. on Machine Learning*, pp.148–156, 1996.

[2] Le Gall JR, Loirat P, Alperovitch A, Glaser P, Granthil C, Mathieu D, Mercier P, Thomas R, Villers D. A simplified acute physiology score for ICU patients. Critical Care Medicine 12(11):975-977, 1984.

[3] Physionet Challenge: http://physionet.org/challenge/2012/

[4] P. Viola and M. J. Jones, "Robust real-time face detection," In *IJCV* 57(2), 137–154, 2004.

*** Corresponding author:**

Yingli Tian
Department of Electrical Engineering
The City College of New York
New York, NY 10031, USA
ytian@ccny.cuny.edu