A Neural Network Model for Mortality Prediction in ICU

Henian Xia[1], Brian J Daley[2], Adam Petrie[3], Xiaopeng Zhao[1]

[1]Department of Mechanical, Aerospace and Biomedical Engineering,
University of Tennessee, Knoxville, USA
[2]Department of Surgery, University of Tennessee Medical Center, Knoxville, USA
[3]Department of Statistics, Operations, and Management Science,
University of Tennessee, Knoxville, USA

Abstract

Scoring the severity of illness of ICU patients can provide evaluation of a patient's situation and thus help doctors make decisions on what treatment to take. This study aimed to develop an artificial neural network model for patient-specific prediction of in-hospital mortality. Data from PhysioNet Challenge 2012 was used. 12,000 records were divided to a training set, a test set and a validation set, each of which contains 4000 records. Outcomes are provided for the training set. A neural network model was developed to predict the risk of in-hospital mortality using various physiological measurements from the ICU. Twenty-six features were selected after a thorough investigation over the different variables and features. A two-layer neural network with fifteen neurons in the hidden layer was used for classification. One hundred voting classifiers were trained and the model's output was the average of the one hundred outputs. A fuzzy threshold was utilized to determine the outcome of each record from the output of the network. Our model yielded an event 1 score of 0.5088 and an event 2 score of 82.211 on the test data set.

1. Introduction

An intensive care unit (ICU) is for patients with the most serious diseases or injuries. Most of the patients need support from equipment like the medical ventilator to maintain normal body functions and need to be constantly and closely monitored. For decades, the number of ICUs has experienced a worldwide increase [1]. During the ICU stay, different physiological parameters are measured and analysed each day. Those parameters are used in scoring systems to gauge the severity of the patients. Many types of severity or prognostic scoring systems have been developed for the ICU, such as the acute physiology and chronic health evaluation system (APACHE II), the simplified acute physiology score (SAPS II) and the mortality probability model (MPM). Those systems are important for many reasons. They provide evaluation of patients' situations so that the intensive care can be restricted to patients most at need. While the intensive care improves the outcome for seriously ill patients, it comes with an expensive cost. In 2005, the mean intensive care unit cost is as high as 31,574±42,570 dollars for patients requiring mechanical ventilation and 12,931 ± 20,569 dollars for those not requiring mechanical ventilation [2]. The mortality assessment is crucial for making the critical decision of whether to interrupt the life-support treatments when intensive care is considered helpless. Besides, the mortality prediction helps doctors determine what treatment process to take.

Most of the prevalent mortality assessment models are developed using linear regression over a score computed from physiological variables. For example, the SAPS II examined 37 variables, and chose 17 that were found to be associated with the hospital mortality most significantly. The 17 variables include 12 physiology variables, age, type of admission (scheduled surgical, unscheduled surgical, or medical), and three underlying disease variables (acquired immunodeficiency syndrome, metastatic cancer, and hematologic malignancy). A score is computed using the 17 variables and is converted to a probability of hospital mortality using a linear regression equation.

More recently, the data mining techniques have been proved to be useful in the ICU mortality prediction [3, 4]. The data mining techniques are used to discover patterns hidden in large clinical data [5]. The volume of clinical data is increasing every single day. It is difficult for human experts to extract information from the data by looking at them manually. In contrast, the data mining techniques can automatically extract information from the raw data [6].

Among the different types of data mining techniques, the artificial neural network is one of the most successful methods. It is widely used because of its capabilities like nonlinear learning, multi-dimensional mapping and noise tolerance [7]. Previous studies reported that the neural network models were better than [3, 4, 8] or at least similar to [9] the linear regression models.

This work is in response to the Computing in Cardiology/Physionet Challenge 2012 "Predicting Mortality of ICU Patients". The focus of the challenge is to "develop methods for patient-specific prediction of in-hospital mortality". In this work, we have developed an artificial neural network model using data collected during the first two days of an ICU stay. The paper is organized as follow: Section 2 gives an introduction to the data available; the feature extraction methods are explained in Section 3; Section 4 gives the prognostic model and the performance metrics; Section 5 presents the results.

2. Data

2.1. Six descriptors and thirty-seven variables

The data consists of 12000 records, each from an ICU stay. The 12000 records are divided to three data sets equally. Four thousand records are used in training set A, and the rest form test sets B and C. The outcomes for the training set are available to us. We developed our algorithms based on the data set A. The data were collected from four types of ICUs: coronary care unit, cardiac surgery recovery unit, medical ICU and surgical ICU. All the ICU stays lasted for at least 48 hours.

Six general descriptors are collected on admission. They are "RecordID (a unique integer for each ICU stay)", "Age (years)", "Gender (0: female, or 1: male)", "Height (cm)", "ICUType", "Weight (kg)". 37 other variables were collected once, more than once, or not at all in each record. In Figure 1, we plotted the numbers of occurrence of each variable in the training set A and the test set B.

2.2. Missing data handling

Up to 42 variables are available. Among them, some variables are very rarely collected. For example, the variables TropI (Troponin-I) and TropT (Troponin-T) never appeared at all. Those variables should be ignored during the feature extraction. Most variables are not collected in every ICU stay. For example, the variable HR (heart rate) is collected in 97% of the records; the variable Lactate is only collected in 55% of the records. For classification purpose, the feature space should be consistent for all the records. Thus it is necessary to handle the missing data properly. For example, for records where the variable Lactate is not collected, we should give the variable Lactate an artificial value.

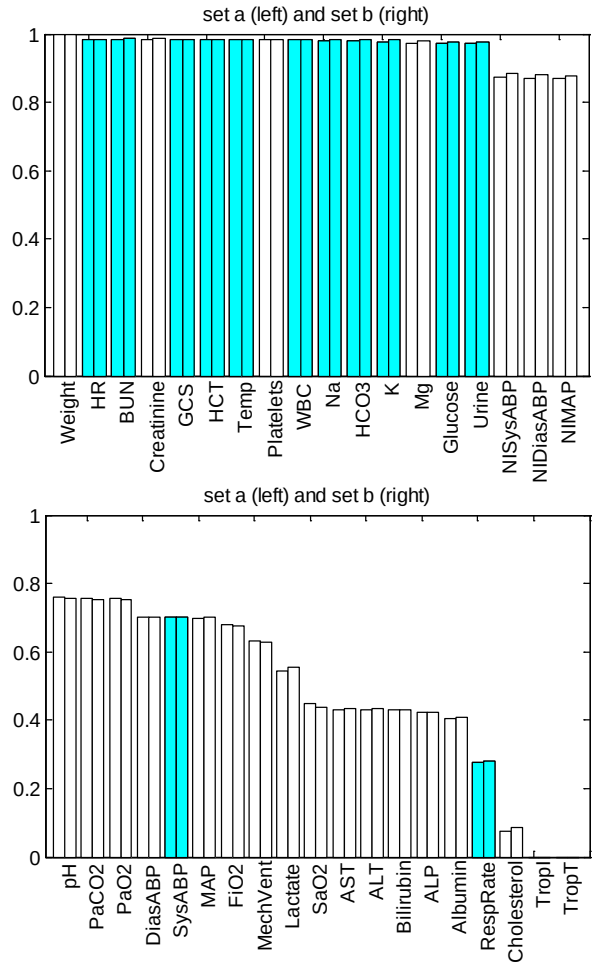

Figure 1. Most frequent 18 variables (top) and least frequent 19 variables (bottom). Variables plotted in cyan are used in the SAPS.

Our way to handle the missing data is based on the assumption that, if one variable is missing, doctors would consider that variable unrelated to the patient's illness. Thus that missing variable should be in normal range. For example, if the variable Temp (Temperature, °C) is missing, it would be in the range of 36 to 38.4.

2.3. Performance evaluation

Two scores were defined to evaluate the algorithms. The first score is the minimum of the sensitivity and the precision. The sensitivity and precision are defined as below:

$$\text{sensitivity} = \frac{TP}{TP+FN}, \text{precision} = \frac{\text{TP}}{\text{TP+FP}}$$

where TP, FP, TN and FN respectively means "true positive", "false positive", "true negative" and "false negative", and are as below:

Table 1. Table of confusion.

Outcome		Observed	
		Death	Survivor
Predicted	Death	TP	FP
	Survivor	FN	TN

The second score is based on a modification of the Hosmer-Lemeshow statistic and the lower this score, the better the algorithm. In this work, we focused our effort on the event 1 score.

3. Feature extraction

After thorough investigation and tests, we have selected 26 features (Table 2). Those features were found to be most distinguishing for the hospital mortality. We computed the mean, minimum, maximum, last data point value and the trend estimation for all variables except the static variable of age, ICU type and gender, resulting in 141 features. The trends of the time series variables are estimated using linear regression. For example, Figure 2 shows the linear regression of the HR in one record, and the trend is estimated as the slope. Inspired by the physiological importance, the sum of urine over the 48 hours is used as a feature. Also the ratio between the FiO2 and PaO2 is used as a feature too.

Table 2. Best 26 features.

Variable	**Feature**	**Event 1 score**
GCS	LastDataPoint	0.375
GCS	WeightedMean	0.33164
GCS	Max	0.31408
HCO3	Min	0.30325
Urine	Sum	0.287
GCS	Slope	0.2852
HCO3	Max	0.26592
BUN	Max	0.26354
BUN	LastDataPoint	0.26354
HCO3	LastDataPoint	0.26256
BUN	Min	0.2599
HCO3	WeightedMean	0.25632
BUN	WeightedMean	0.2471
SysABP	WeightedMean	0.24549
WBC	LastDataPoint	0.24368
SysABP	LastDataPoint	0.23944
FiO2, PaO2	Ratio	0.22754
WBC	WeightedMean	0.22744
Temp	WeightedMean	0.21775
Glucose	Max	0.21732
Na	WeightedMean	0.21525
Na	Max	0.21342
SysABPNISysABP	Min	0.21261
Age		0.2112
Lactate	LastDataPoint	0.21119
Temp	LastDataPoint	0.21078

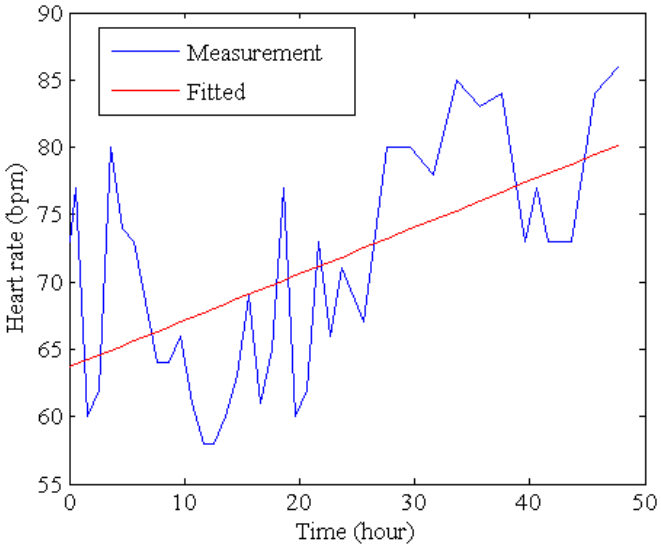

Figure 2. Trend estimation: the measurements were fitted to a linear line, and the slope of the line was used for the trend estimation

Each feature was independently used for the mortality prediction and the performance for each feature was evaluated using the event 1 score. The best individual feature was identified as the last data point value of the GCS which yielded an event 1 score of 0.375. A list of the best individual features is shown in Table 2. After finding the best individual features, we investigated the performances of different combinations. It was found that, the combination of the best 26 features gave the best performances in the event 1 score.

4. Classification

4.1. Neural network

During neural network training, the optimization could often be stuck in local minima and result in very poor classification accuracy. In this work, we used a "voting" strategy to overcome that problem. For each training set, we repeatedly train 100 neural networks; in prediction, the 100 neural networks each predict an intermediate probability for an input, and the final output is the average of the 100 intermediate probabilities.

4.2. Oversampling

The training data set contains only 13.85% of its records as dead (positive) records. To expose the positive records more frequently to the training algorithm, we tried oversampling the positive records. That means we may input the same positive records to the neural network during training more than once.

4.3. Fuzzy threshold

The output of the neural network is a score between 0 and 1. In this problem, the best threshold to differentiate the negative and positive records is typically smaller than 0.5, and may fluctuate because of the number of voting classifiers used, the number of positive records oversampled, etc. During training, we are able to find out the optimal threshold, but the same threshold may not be optimal for test data set. To overcome this, we used a "fuzzy threshold" for test. For example, if it was found that the best threshold in training was 0.35, during test, we would determine all records with a score below 0.34 as negative, all records with a score above 0.36 as positive, and would randomly guess the records with a score between 0.34 and 0.36. This approach was able to give a performance close to the optimal threshold.

5. Results

The algorithms were trained and tested in the training set using the 5-fold cross validation. Each test was repeated for 5 trails to reduce the influence of the randomness. Different parameter settings were investigated thoroughly in order to find the optimal model.

The following network architectures were tested: one-layer neural network with 3 to 20 hidden neurons and two-layer network with 3 to 7 neurons in the first and second hidden layers respectively. It turned out that, a two-layer artificial neural network with fifteen neurons in the hidden layer gave the best performance. The hyperbolic tangent sigmoid transfer function was used in each layer. Also different training functions like the Bayesian regulation back-propagation, the Conjugate gradient back-propagation, Levenberg-Marquardt back-propagation and so on. The Levenberg-Marquard back-propagation algorithm was found to be the best.

Also the number of positive records to oversample was adjusted. The final model oversampled 70% of the positive records and used a fuzzy threshold band of (0.35, 0.37). On the training set, it gave an event 1 score of about 0.495 and an event 2 score of about 57 in 5-fold cross validation. On the test data set, an event 1 score of 0.5088 and an event 2 score of 82.211 were obtained.

Acknowledgements

This work was in part supported by the NSF under grant number CMMI-0845753.

References

[1] Hanson C, Marshall B. Artificial intelligence applications in the intensive care unit. Crit Care Med 2001;29(2):1-9

[2] Dasta JF, McLaughlin TP, Mody SH, Piech CT. 2005 Daily cost of an intensive care unit day: the contribution of mechanical ventilation. Crit Care Med 2005;33(6):1266-71.

[3] Dybowski R,Weller P, Chang R, Gant V. Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. Lancet 1996;347(9009): 1146-50.

[4] Nimgaonkar A, Sudarshan S. Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. Intensive Care Med 2004;30:248-53.

[5] Cios K, Moore G. Uniqueness of medical data mining. Artif Intell Med 2002;26:1-24.

[6] Hand D, Mannila H, Smyth P. Principles of data mining. Cambridge, MA, USA: MIT Press, 2001.

[7] Haykin S. Neural networks - a compreensive foundation, 2nd ed, New Jersey, USA: Prentice-Hall, 1999.

[8] Silva A, Cortez P, Santos MF, Gomes L, Neves J 2006 Mortality assessment in intensive care units via adverse events using artificial neural networks Artificial Intelligence in Medicine 36:223-234

[9] Wong L, Young J. A comparison of ICU mortality prediction using the APACHE II scoring system and artificial neural networks. Anaesthesia 1999;54:1048-54.

Address for correspondence.

Xiaopeng Zhao
Department of Mechanical, Aerospace, and Biomedical Engineering
University of Tennessee
Knoxville, TN 37996-2210
USA
xzhao9@utk.edu