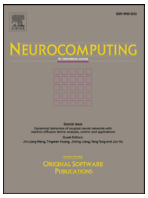
Neurocomputing 000 (2017) 1–8

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Mortality prediction for ICU patients combining just-in-time learning and extreme learning machine

Yangyang Ding[a], Youqing Wang[a,b,*], Donghua Zhou[b]

[a] College of Information Science and Technology, Beijing University of Chemical Technology, China
[b] College of Electrical Engineering and Automation, Shandong University of Science and Technology, 579# QianWanGang Road, Qingdao 266590, China

ARTICLE INFO

Article history:
Received 27 June 2017
Revised 27 August 2017
Accepted 28 October 2017
Available online xxx

Communicated by Dr. Ma Lifeng Ma

Keywords:
Extreme learning machine (ELM)
Just-in-time learning (JITL)
Intensive care unit (ICU)
Mortality prediction
Patient-specific model

ABSTRACT

Mortality prediction for patients in intensive care unit (ICU) is necessary to prioritize resources as well as to help the medical staff to make decisions, and hence more accurate methods for identifying high risk patients are very important for improving clinical care. However, many existing approaches including some scoring systems now being used in the hospital are not good enough since they try to establish a global/average offline model, which may be unsuitable for a specific patient. Thus, a more robust and effective monitoring model adaptable to individual patients is needed. To establish a more personalized model, this study proposes a two-step framework, in which the first step is for clustering and while the second one is for mortality predication. A novel method combining just-in-time learning (JITL) and extreme learning machine (ELM), referred to JITL-ELM, is proposed for mortality prediction, which applies global optimization of variables and neighborhood of appropriate samples to build an accurate patient-specific model. In addition, a simplified JITL-ELM with less key physiological variables is developed. In the experiment, 4000 real clinical records of ICU patients are collected to validate the proposed algorithm, of which the AUC index is 0.8568, which is much better than the existing traditional global/average models, and furthermore the simplified JITL-ELM still performs well.

1. Introduction

Intensive care unit (ICU) admits only the most severely ill patients who require life-sustaining treatments or extensive monitoring. Inside ICU, the most advanced monitoring equipment and emergency facilities of the hospital are centralized, which makes it playing an important role in enhancing the success rate of emergency treatment and further reducing the mortality. Mortality prediction in ICU can reflect the severity of disease or the prognosis of patients, get a more reasonable allocation of medical resources and helps clinicians to make decisions. Hence, mortality prediction for ICU patients is always one of the most important topics in clinical and healthcare research, which causes a wide concern of researchers.

Popular methods belonging to generalized linear models are commonly used in mortality prediction, in the form of many prognostic scoring systems. Among which three scoring methods are mostly used, i.e., Acute Physiology and Chronic Health Evaluation (APACHE) scoring system [1], Simplified Acute Physiology Score (SAPS) [2], and mortality probability model (MPM) [3]. For the latest version of these systems, the worst physiological values in the first 24 h after patients entered ICU are used to establish the logistic regression (LR) model, whereas the other data are not used, and this leads to the loss of information. In addition, linear model makes it inaccurate to predict the patients' status as well.

With deeper research and development of technology, as well as based on the increasing volume of clinical data in ICU, more and more researchers prefer to use data-driven learning approaches for mortality prediction. Machine learning, which belongs to nonlinear modeling method, such as artificial neural networks [4,5], support vector machine [6,7], decision tree [8], naive Bayesian model [9,10], as well as more complex models, such as incremental information network [11], have been explored to portray patients' characteristics in the past decades, and they also give more accurate results only depends on numbers of physiological measurements.

All these above-mentioned methods tried to use a global predictive model derived from all available training data to compute risk scores for a query patient, however, various patients behave in highly individual ways. Over the last decades, different patient-specific models have been developed for decision support [12] or adaptive monitoring in critical care [13,14]. A personalized model

[*] Corresponding author at: College of Electrical Engineering and Automation, Shandong University of Science and Technology, 579# QianWanGang Road, Qingdao 266590, China.
E-mail address: wang.youqing@ieee.org (Y. Wang).

https://doi.org/10.1016/j.neucom.2017.10.044
0925-2312/© 2017 Elsevier B.V. All rights reserved.

in the context of healthcare applications has recently been investigated, among which a personalized modeling method that leverages evolutionary optimization techniques is proposed in [15], which is used in some specific fields such as personalized drug design. Moreover, just-in-time learning (JITL) and principal component analysis (PCA), referred to learning-type PCA (L-PCA) [16], was combined to build an online individual-type model to monitor the patient's status, in which JITL gathers the most relevant samples for adaptive modeling of complex physiological processes, and PCA was used for personalized modeling. Recently, another novel personalized modeling approach named JITL-ELM, which integrates JITL and extreme learning machine (ELM), was proposed [17]. JITL borrows the diagnostic idea of "similar symptoms characterize similar results" by searching for the most relevant samples to establish a patient-specific model, aiming at improving the precision, while ELM was chosen for mortality prediction. Based on these studies, the newly proposed method builds a personalized model that is more suitable for the query patient using data collected from other patients, which is also the main difference from traditional models. In brief, the study aims to make some extensions and increase the value of its clinical application by developing the personalized model.

To evaluate different algorithms, area under the receiver-operating curve (AUC) is used in the experiment, and algorithms with AUC closer to 1.0 have better classification performance or higher diagnostic value. With real physiological data of ICU patients, the AUC of JITL-ELM is 0.8568, with sensitivity of 0.7655 and specificity of 0.7907. Compared with ELM, Back Propagation (BP) neural network, LR, and SAPS-I, JITL-ELM gains 3.53%, 8.67%, 22.12%, and 25.69% increases of the AUC index, respectively. In order to improve the calculation speed and to narrow the search scope, a preprocess work for clustering needs to be conducted before prediction, which leads to the born of a two-step framework in this study. In addition, the study tries to establish a more practical model with a few key physiological variables, which is also performs much better than the SAPS-I system.

In summary, the main findings and contributions are as follows. (1) This study provides several improvements on the method in [17], which performs the best for improving mortality prediction accuracy compared with other conventional methods. (2) A two-step framework is constructed, in which the first step is used for clustering, whereas the second step for patient-specific mortality prediction. With the virtue of speeding up the calculation, the framework makes JITL-ELM algorithm more practical and have more advantages in clinical promotion. (3) Experiments show that a small number of key variables can also achieve a better mortality prediction, which is superior to the other typical methods especially SAPS-I. Moreover, JITL can solve the practical binary classification problem with unbalanced distribution to some extent, and the idea of "similar input produces the similar output" makes it highly descriptive.

The remainder of this paper is organized as follows. Section 2 outlines the related work including ELM, JITL, their combination JITL-ELM and its promotion, as well as some related evaluation metrics. Some introductions about data sources and pretreatment before the experiment are provided in Section 3. Then the experimental results and discussion are presented in Section 4. Finally, some conclusions are drawn in Section 5.

2. Related work

2.1. Extreme learning machine

ELM is one of the leading trends for fast learning, which was proposed in literature [18,19], of which a brief introduction is conducted as follows.

Given the dataset $(X_i, y_i)_{i=1}^{N}$, where $X_i \in R^{1\times m}$ indicates an input training sample, and y_i is a scalar, which represents the label of categories. Then ELM model can be described as

$$\sum\nolimits_{l=1}^{L} \beta_l g_l (w_l X_i^T + b_l) = y_i, i = 1,\ 2, \ldots,\ N \tag{1}$$

where L denotes the number of hidden nodes, $w = (w_1^T, w_2^T, \ldots, w_L^T) \in R^{m\times L}$ indicates the weight vectors between the input and hidden layers; $\beta = (\beta_1^T, \beta_2^T, \ldots, \beta_L^T) \in R^{L\times t}$ is the weight vector between the hidden and output layers; $g(\cdot)$ is the activation function, t indicates the number of output layer.

By generating the weight matrix $w = (w_1, w_2, \ldots, w_L)^T$ and offset vectors $b = (b_1, b_2 \cdots b_L)^T \in R^{L\times 1}$ randomly, the output matrix H of hidden layer can be computed as

$$H(w_1, w_2, \ldots, w_L; b_1, b_2 \ldots b_L; X_1, X_2, \ldots, X_N)$$
$$= \begin{bmatrix} g(w_1X_1^T + b_1) & g(w_2X_1^T + b_2) & \cdots & g(w_LX_1^T + b_L) \\ g(w_1X_2^T + b_1) & g(w_2X_2^T + b_2) & \cdots & g(w_LX_2^T + b_L) \\ \vdots & \vdots & \ldots & \vdots \\ g(w_1X_N^T + b_1) & g(w_2X_N^T + b_2) & \cdots & g(w_LX_N^T + b_L) \end{bmatrix}_{N\times L} \tag{2}$$

It is clear that the only unknown variable is β, and according to the mathematical model of the single hidden feedforward networks (SLFNs) given by

$$H\beta = Y \tag{3}$$

of which the least square solution with minimal norm is analytically determined using Moore-Penrose generalized inverse H† [20]:

$$\beta = H^{\dagger}Y \tag{4}$$

To obtain a better generalization performance [21], a regularization parameter C is often added into (4), expressed as

$$\beta = \begin{cases} H^T\left(\frac{I}{C} + HH^T\right)^{-1}Y \ when\ N < L, \\ \left(\frac{I}{C} + H^TH\right)^{-1}H^TY \ when\ N > L, \end{cases} \tag{5}$$

and there are two options for users according to the size of training data.

Unlike the other traditional learning algorithms, like BP or SVM, the parameters of hidden layers of ELM are randomly assigned and prefixed, and Huang et al. have proved that SLFNs with randomly generated hidden neurons and output weights computed by ridge regression still maintain the universal approximation capability of SLFNs [22], which improves the training speed greatly.

Through the introductions above, it can be concluded that ELM stands out from the other neural network methods with the following virtues: extremely fast training speed, good generalization, as well as the universal approximation capability.

2.2. Just-in-time learning

Just-in-time learning (JITL) algorithm has been widely applied for system identification and online soft sensing in chemical processes, but rarely applied in medical field. However, its idea of "similar inputs produce similar output" is closely related to the procedure of diagnosis disease for patients. Doctors often come to conclusions based on similar cases that have been diagnosed, which is also the gist of our algorithm. The mechanism of JITL is introduced as follows.

Considering the same dataset $(X_i, y_i)_{i=1}^{N}$ in the previous section, a conventional model tries to build a fine mapping relationship $f(\cdot,\ \cdot)$ between the input and output data written as

Please cite this article as: Y. Ding et al., Mortality prediction for ICU patients combining just-in-time learning and extreme learning machine, Neurocomputing (2017), https://doi.org/10.1016/j.neucom.2017.10.044

$y_i = f(X_i, \theta) + \varepsilon_i$, and then it is converted into solving the following optimization problem:

$$\theta^* = \arg\min_{\theta} \sum_{(X_i, y_i)} (y_i - f(X_i, \theta))^2 \tag{6}$$

where θ indicates the coefficient vector and ε_i is the modeling error which satisfies a Gaussian distribution.

Differently, JITL tends to establish a local model in the neighborhood space of each query sample by collecting the corresponding similar dataset, which can be expressed as

$$\theta^* = \arg\min_{\theta} \sum_{(X_i, y_i) \in \Omega} (y_i - f(X_i, \theta))^2 \mu_i \tag{7}$$

$$\mu_i = \exp\left(\frac{-\|X_i - X_q\|^2}{2\sigma^2} \right) \tag{8}$$

where Ω indicates the domain space composed of its k relevant samples, μ_i denotes the weight between the training data and the query data X_q according to Gaussian Kernel Function refer to (8), which reflects the influence on the result that similar samples exert; and σ reflects the width parameters of the kernel function.

The domain space Ω can be determined by several approaches, and a brief method using synthetic distance $d(\cdot, \cdot)$ [23], which integrates the Euclidean distance $E(\cdot, \cdot)$ and Angle distance $\cos(\cdot, \cdot)$, is adapted in this study, which can be written as

$$d(X_q, X_i) = \lambda \sqrt{e^{-E(X_q, X_i)^2}} + (1 - \lambda) \cos(X_q, X_i) \tag{9}$$

where $E(X_q, X_i) = \sqrt{(X_q - X_i)^T (X_q - X_i)}$, $\cos(X_q, X_i) = \frac{{X_q}^T X_i}{\|X_i\|_2 \|X_i\|_2}$, and $\lambda \in (0, 1)$ indicates the weight coefficient.

Then Ω can be described as

$$\Omega = \left\{ X_q^1, X_q^2 \ldots, X_q^k \right\} = \{ X_i | d(X_i, X_q) < h \} \tag{10}$$

where h indicates the radius of domain space, determining the size of the similar data set; $X_q^1, X_q^2 \ldots, X_q^k$ indicates the most similar k samples to the query data X_q, which are inside the domain space indicated by Ω. However, the value of h is hard to determine in practical application, for which some prior knowledge or a certain range of search is needed. The computational difficulty of the algorithm will be increased if h is too large while the generalization ability will decline if h is too small. Therefore, the number of similar samples k is chosen to determine the size of domain space for convenience.

In conclusion, JITL collects the relevant samples instead of the whole dataset as the training samples to establish an online model for each query sample, of which the establishment and prediction process is conducted locally, which leads to a better online adaptive capability compared to the conventional modeling methods. Moreover, the collection of similar data contributes a lot to solve the problem of imbalanced class distribution.

In order to illustrate the differences between JITL and the conventional modeling mechanisms more clearly, their flow diagrams are given in Fig. 1. The global modeling method uses the whole data as training samples to build an offline model, of which the structure is fixed. However, JITL collects the similar samples to the query data for online modeling, in which the structure and parameters are variable.

When a new patient comes, doctor will restart to diagnosis of disease by searching for the similar cases that he has known. Similarly to the diagnosis process, a local model will be established newly when a new query test sample is given, which also reflects the idea “similar illness produces similar symptoms most likely”.

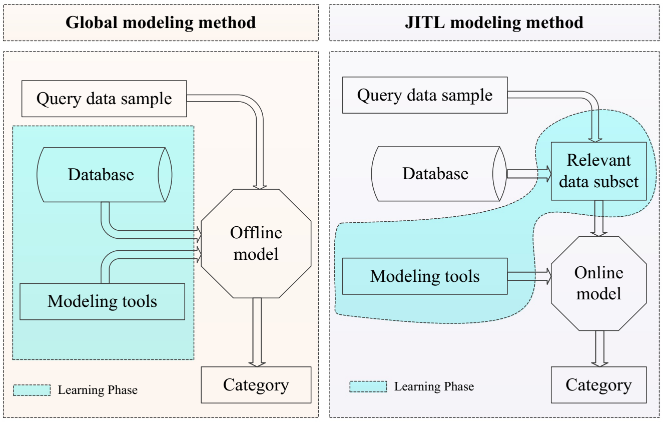


Fig. 1. Mechanism comparison of conventional model and JITL-based model.

2.3. JITL-ELM

Based on the idea of “similar inputs produce similar output”, three steps including “relevant dataset selection”, “local model construction” as well as “output forecast” will be followed in order to build a personalized model for each query data sample. In this study, ELM is selected as the locally modeling tool which is mentioned in Fig. 1 to achieve the purpose of classification. To make it clearer, the pseudo code is provided in Table 1 to describe the entirely procedure of personalized modeling for the current query patient.

In fact, the total number of patients used in this study is only 4000, so “leave one out cross” method is utilized in order to make full use of the existing data. In other words, patients available will be regarded as a test/query sample successively, meanwhile the remaining 3999 ones constitute the historical database, from which some samples will be chosen for further modeling. It is a measure to ensure that the samples in the historical database are adequate for use, while in practical applications, there is no need to use “leave one out cross” method with sufficient data samples. Moreover, if the related dataset belongs to the same category, ELM cannot work properly, and some measures must be taken. As described in pseudo code, actually when one of the categories has too much samples, namely more than 90% in related dataset, samples of the minority category will be randomly selected from historical database, and then they are added into related dataset.

It can be seen that the predicted result is estimated only based on the selected subset of samples, which makes the principle of similarity measurement and the selection of similar subset play a crucial role in the JITL strategy. There are also a lot of other methods to determine the subset.

In terms of similarity measurement, Fujiwara et al. confirmed the relevant subset based on the Q and T^2 statistics after PCA projection [24], but it is difficult to be applied in process online modeling due to its great computational cost. Chen et al. [25] proposed a new similarity measurement utilizing a new SLPP version for regression problems to select relevant samples and determine the weights of relevant features, which leads to a high precision under low computation complexity. The method adopted in this study uses synthetic distance instead of judging Euclidean distance and angle distance, respectively, which is effective and simple both in calculation and dissemination.

In terms of the relevant subset selection, the method of “increasing one by one” is also a good candidate by judging the issue whether the performance of the model is improved after the addition of a new sample, which evidently leading to a heavy computation as well as the possible problem declining the final

Please cite this article as: Y. Ding et al., Mortality prediction for ICU patients combining just-in-time learning and extreme learning machine, Neurocomputing (2017), https://doi.org/10.1016/j.neucom.2017.10.044

Table 1
Pseudo code of JITL-ELM algorithm.

```
Procedures of JITL-ELM:
Input:
•  history_data: The class-labeled data set in the historical database ($T$ samples)
•  $X_q$: the current query patient data
•  num_JITL: the number of similar samples belonging to the domain space
Output:
•  predicted category of query/test sample: query_result
for the query patient $X_q$
  relevant dataset: relavant_data=JITL(history_data, $X_q$)
    training dataset: train_data=relavant_data


    query_result=ELM(train_data, $X_q$)

end for
Procedure of JITL (namely JITL function)
Input:
•  query data: $X_q$
•  historical dataset: history_data
Output:
•  relevant dataset: relavant_data
For $i = 1$ to $T$
    Compute the synthetic distance:$\bar{d} = \begin{cases} d(X_q, X_i), & if\,\cos(X_q, X_i) \geq 0 \\ 0, & if\,\cos(X_q, X_i) < 0 \end{cases}$
    Sort the $N - 1$ samples in descending order according to $\bar{d}$.
    relavant_data=samples ranking in the top num_JITL.
End for
If 90% relavant_data belong to the same category:
    a)  num_sample_select=majority class samples number – minority ones
    b)  Randomly select num_sample_select samples in minority class category, then add them into relavant_data
End if
Procedure of ELM (namely ELM function)
Input:
•  training dataset: train_data
•  category of the training dataset: Y
•  query data: $X_q$
Output:
•  predicted category of the testing dataset: query_result
(a) Randomly generated the weight coefficient $w$ and bias vector $b$ of the input layer
(b) Compute the output of hidden layer $H(w, b, train\_data)$ according to Eq. (2)
(c) Calculate weight coefficient $\beta$ between the hidden and output layer, according to Eq. (5)
(d) Predict the category of test/query sample: $test\_result = H(w, b, X_q)\beta$
```

performance caused by very small subset size [26]. In the practical application, the similar data sets can be determined by setting weights according to (8) or the threshold of synthetic distance, because there are a wide range of cases available to be chosen in the actual hospital case library. To guarantee the adequate training samples, the relevant subset is confirmed by determining the proper number of similar samples k.

2.4. Two-step JITL-ELM

JITL aims to obtain the prediction result of the current testing patient with the k most relevant samples collected from the historical database, and the forecast is operated for patients one by one, namely the model changes for each patient. However, this search mode has many problems, such as time-consuming and large calculation.

In order to overcome the problems described above, clustering analysis is conducted to fractionize the samples firstly, of which two approaches are offered in the study, that is, “ICU-type clustering approach” and “Ward’s clustering approach”.

In terms of “ICU-type clustering approach”, patients are divided into clusters according to known ICU type, in which the prior knowledge is utilized; while for “Ward’s clustering approach”, which emphasizes the internal differences of similar samples should be small (namely the variance or standard deviation) to ensure a large degree of similarity within the same cluster, but little similarity between different clusters, which highlights the homogeneity of the same area. Furthermore, Ward’s method (also called the sum of squared deviation method) [27], which measures the distance between the two clusters by squared Euclidean distance. At the initial step, all clusters are singletons (clusters containing a single sample), then the centroid variance will be calculated so that the clusters with the increase of the minimum will be merged preferentially, finally the other clusters can be merged in turn gradually, as desired.

Hence, the mortality prediction algorithm for ICU patients is divided into two steps, referred to “two-step mortality prediction” algorithm in this study, of which the procedure is illustrated in Fig. 2. The first step, referred to the clustering step, aims to cluster the data samples, while the second one, called classification step, tries to realize the purpose of mortality prediction using JITL-ELM approach. The addition of clustering step aims to narrow the search scope and improve the retrieval speed.

2.5. Evaluation metrics

To measure the effectiveness of a classifier, several indices are used in this section. The traditional index, overall accuracy, is no longer applied to the dataset with imbalanced problem, because it has a natural tendency to favor the majority class by assuming balanced class distribution or equal misclassification cost. For mortality prediction in this study, imbalanced problem also exists with small amounts of dead patients and majority of survivals.

Please cite this article as: Y. Ding et al., Mortality prediction for ICU patients combining just-in-time learning and extreme learning machine, Neurocomputing (2017), https://doi.org/10.1016/j.neucom.2017.10.044

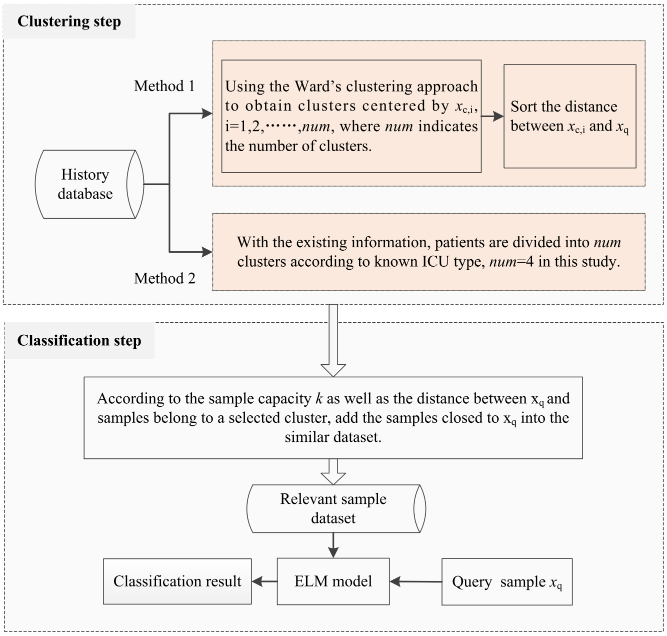

Clustering step
Method 1
Using the Ward's clustering approach to obtain clusters centered by $x_{c,i}$, i=1,2,⋯⋯,num, where num indicates the number of clusters.
Sort the distance between $x_{c,i}$ and x_q
History database
With the existing information, patients are divided into num clusters according to known ICU type, num=4 in this study.
Method 2
Classification step
According to the sample capacity k as well as the distance between x_q and samples belong to a selected cluster, add the samples closed to x_q into the similar dataset.
Relevant sample dataset
Classification result
ELM model
Query sample x_q

Fig. 2. Procedure of two-step JITL-ELM.

Table 2
Confusion matrix.

		Real status	
		Positive	Negative
Predicted status	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

In the field of medicine, people prefer to use “positive” and “negative” to represent two categories, the former indicates the dead class, while the latter means the survival ones.

Confusion matrix is a commonly used method to evaluate the accuracy of classification result, as shown in Table 2, and TP, FP, FN, TN indicate the number of samples of “True Positive”, “False Positive”, “False Negative” and “True Negative” in the experiment, respectively. Furthermore, there are lots of the other indices, such as,

$$sensitivity = \frac{TP}{TP+FN}$$

$$specificity = \frac{TN}{TN+FP}$$

$$G\text{-}mean = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}}$$

$$FPR = \frac{FP}{TN+FP} = 1 - specificity$$

where sensitivity (also called Recall, or True Positive Rate, shorted for TPR) means the proportions of real positive samples can be detected correctly, and specificity indicates the rate of real negative samples are not wrongly misclassified; G-mean, referred to the geometric mean of those accuracies, indicates the final result to measure the functionality of a classifier; False Positive Rate (FPR) denotes the rate of real negative samples that are misclassified.

Another useful tool, receiver-operating curves (ROC) graph [28], provides a visual illustration of the performance of classifiers on binary datasets, where a classifier corresponds to a point, x-coordinate represents FPR, and y-coordinate denotes TPR, which leads to classification results for both the two classes are perceivable with a single point. That is to say, the performance exhibited by ROC graph is independent of the class distribution and cost information.

The classification is carried out by setting a proper threshold, then the predictive result values greater than the threshold are set to 1 (1 indicates death or positive), while the others are 0 (0 indicates survival or negative). Different thresholds will lead to different classification performance, including the specificity and sensitivity, and then ROC curve is born. All the researchers need to do is to find a proper tradeoff between the ability to identify the positive samples and the negative ones. In this study, the appropriate threshold is selected by finding out the best G-mean values, which can ensure that both of the two categories have a good classification accuracy.

Moreover, a derived index called AUC, referring to the area under the ROC curve, is often used to evaluate the performance of a binary classifier quantitatively. The closer AUC index to 1.0, the better the classification results.

3. Data sources and processing

3.1. Data sources

In order to validate the performance of the proposed algorithm, physiological data of ICU patients are collected from a website named PhysioNet [29], which offers free access to complex physiological signals and biomedical signal research resources, and its scientific and rigorous have been widely validated, possessing a high authority.

In this study, 4000 records of ICU patients are collected totally, including 554 dead patient and 3446 survival ones, of which the average age is 64.25 years, and men accounts for 56.2% of the proportion. The largest number of patients was admitted to the medical ICU (35.8%), followed by the surgical (28.4%), cardiac surgery recovery (21.1%), and coronary (21.1%) ICUs.

For each ICU patient, the data collected from the first 48 h of ICU stay are generally composed of three parts, including the basic information of the hospital admission (RecordID, age, height, weight, and ICU type), the time series measurements for 36 physiological variables, and the final state (0 = survivor, 1 = died). Taking two patients for example, the initial ICU data sets are shown in Table 3. It is easy to see that,

- For the same patient, the sampling frequency is not fixed for the same physiological variable.
- For the same patient, the sampling frequency of different physiological variables is different.
- For different patients, only part of physiological data is collected at the same sampling point.
- The data contain error values, which is not displayed in the table.

Furthermore, part of physiological parameters is illustrated in Table 4, including their abbreviations and full names.

3.2. Data preprocessing

Three issues need to be solved according to the analysis above.

- Selection of effective physiological variables.
- The method to extract the physiological information for each patient.
- Elimination of error values and imputation for missing values.

First, according to medical information and the missing status of each physiological variables, 24 physiological parameters recorded for more than 75% patients are selected, which are shown in Table 4.

Table 3
Physiological dataset of two sample patients.

RecordID = 132539					
Time	HR	Temp	GCS	...	NIDiasABP
0:00	−1	−1	−1	...	−1
0:07	73	35.1	15	...	65
0:37	77	35.6	−1	...	58
1:37	60	−1	−1	...	62
2:37	62	−1	−1	...	52
3:08	−1	−1	−1	...	−1
...	...	...	...	...	...
46:37	−1	−1	−1	...	−1
47:37	86	37.8	15	...	128
RecordID = 132540					
Time	HR	Temp	GCS	...	NIDiasABP
0:00	−1	−1	−1	...	−1
0:42	−1	−1	−1	...	−1
1:11	88	35.2	−1	...	−1
1:26	88	35.1	3	...	−1
1:27	−1	−1	−1	...	−1
1:31	88	34.8	−1	...	−1
...	...	...	...	...	...
46:15	−1	−1	−1	...	−1
47:11	−1	37.1	15	...	49

"−1" denotes the missing value.

Second, the summary statistics (minimum, maximum, mean, standard deviation, 1/4 site, and 3/4 site), age and body mass index (BMI), as well as the indices with physiological meanings are collected for each patient, and 147 features are obtained in final, which will be regarded as the input of model.

Finally, Chauvenet criterion [30] is used to delete the error values, but for the missing data, the median values are selected as the interpolation data for each physiological variables according to the ICU type and age stages. PCA and its variants are often used for data reduction or monitoring in industrial process [31]. In the study, PCA is adopted as pre-processing methods to remove noise and to reduce computational complexity by reducing the dimensions before features are fed into the model. Eventually, the dimension of data is reduced from the original 147 to 46.

4. Results and discussion

4.1. Results using different methods

In this section, the performance of JITL-ELM is compared with some typical algorithms, i.e., ELM, LR, BP neural network. In addition, another reference results of SAPS-I scoring system is also plotted, because it is often used as a kind of indicator to patients' status in hospital.

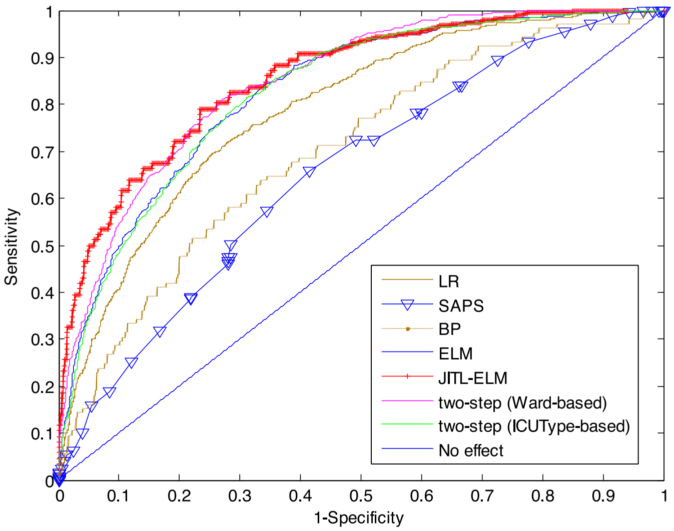


Fig. 3. ROC curves for ELM, BP, LR, JITL-ELM, two-step strategy methods, and SAPS-I scoring system. 400 similar data was selected for the current testing data when using JITL-ELM model here.

Performance results in terms of ROC when using different methods described above, as well as JITL-ELM based on two-step framework are shown in Fig. 3. As a contrast curve, the diagonal line indicates a useless classifier judging death randomly. Generally, if the curve is closer to the coordinate point (0, 1), the classifier's performance will be better. It is clear that the result is improved after adding JITL part, compared with the pure ELM model. Furthermore, JITL-ELM performs best among all these methods, and over-fitting occurred when using BP algorithm, which decreased its accuracy. Another evaluation index called AUC, which denotes the area under the ROC curve is utilized for quantitative evaluation, and the results of classification using the above-mentioned algorithms are shown in Table 5. As far as the AUC index is concerned, the performance of ELM shows a 3.53% percent increase due to the combination of JITL, and results of JITL-ELM reveal an improvement of 8.67%, 22.12% and 25.69% compared to LR, BP and SAPS-I, respectively.

In addition, although the performance of two-step JITL-ELM slightly decreases, it still superior to the others. In terms of the clustering scheme, ward's method performs better.

According to the results in [17], although the AUC value can be optimized with the increasing k, however, the degree of optimization is very weak, and it is at the expense of reducing sensitivity. Through trial and explore, k is set by 400 in this study when using the JITL-ELM model. Moreover, a grid search

Table 4
List of selected physiological parameters and their abbreviations.

Abbreviation	Name	Abbreviation	Name
HR	Heart rate (bpm)^A	Glucose	Serum glucose (mg/dL)
BUN	Blood urea nitrogen (mg/dL)	K	Serum potassium (mEq/L)
GCS	Glasgow coma index ⟨3–15⟩^{B}	Mg	Serum magnesium (mmol/L)
Creatinine	Serum creatinine (mg/dL)	Na	Serum sodium (mEq/L)
DiasABP	Invasive diastolic arterial blood pressure (mmHg)	MAP	Invasive mean arterial blood pressure (mmHg)
NIMAP	Non-invasive mean arterial blood pressure (mmHg)	SysABP	Invasive systolic arterial blood pressure (mmHg)
FiO$_2$	Fractional inspired O$_2$ ⟨0–1⟩	HCO$_3$	Serum bicarbonate (mmol/L)
PaO$_2$	Partial pressure of arterial O$_2$(mmHg)	PaCO$_2$	Partial pressure carbon dioxide
NIDiasABP	Non-invasive diastolic arterial blood pressure (mmHg)	NISysABP	Non-invasive systolic arterial blood pressure (mmHg)
HCT	Hematocrit (%)	Temp	Temperature (°C)
pH	Arterial pH ⟨0–14⟩	Urine	Urine output (mL)
Platelets	Blood platelet (cells/nL)	WBC	White blood cell count(cells/nL)

A: Physical unit.
B: Range of normal values.

Table 5
Classification performance results using different modeling methods.

	LR	SAPS	BP	ELM	JITL-ELM	Two-step (ward)	Two-step (ICUType)
AUC	0.7883	0.6817	0.7016	0.8276	**0.8568**	0.8477	0.8258
Maximum *G*-mean	0.7204	0.6203	0.6520	0.7489	**0.7780**	0.7689	0.7498
Sensitivity	0.7449	0.5856	0.6662	0.7327	**0.7638**	0.7278	0.7194
Specificity	0.6968	0.6570	0.6381	0.7653	0.7907	**0.8123**	0.7816
Accuracy	0.7034	0.6471	0.6420	0.7608	0.7872	**0.8006**	0.7729

Table 6
Classification results using JITL-ELM and simplified JITL-ELM.

	JITL-ELM	Simplified JITL-ELM
AUC	0.8568	0.8278
Max_Gmean	0.7780	0.7472
Sensitivity	0.7655	0.7014
Specificity	0.7907	0.7960
Accuracy	0.7872	0.7829

of $C\{2^{-18}, 2^{-16}, \ldots, 2^{48}, 2^{50}\}$ and the number of hidden nodes $L\{10, 20, \ldots, 90, 100\}$ is conducted in seek of the optimal result, and finally $C = 2^{-10}$ and $L = 25$ are chosen in the experiment, and the weight coefficient λ is selected by 0.7.

Table 5 reports the final sensitivity and specificity for the maximum *G*-mean value. The results are significantly improved after adding the JITL strategy. Moreover, the higher specificity of two-step JITL-ELM (ward-based) denotes it has a better ability to identify the negative samples among the unknown sample set, but its sensitivity is not as ideal as JITL-ELM. That is to say, it improves the specificity at the expense of sensitivity, and that is also the reason that its traditional evaluation index, accuracy, better than others. Researchers can adjust the suitable threshold to achieve a nice tradeoff to gain a satisfying sensitivity and specificity simultaneously.

What worth mentioning, the traditional models uses the whole dataset for modeling, which ignores the specific information of the current query patient. By contrast, JITL collected the similar samples to establish patient-specific model for each patient, and it can also solve the low prediction accuracy problem caused by the distribution imbalance of training samples, which helps to build a more accurate local model. In summary, JITL-ELM algorithm is a good candidate to establish a patient-specific model as well as to promote the accuracy of the mortality prediction.

4.2. *JITL-ELM results after deleting some physiological variables*

In the experiment, 24 physiological variables are selected for modeling and finally gains a good classification result. However, some of the physiological variables may be not measured in the actual monitoring process, which will cause a poor performance like the scoring criterion, such as the APACHE system. Hence, the study tries to cut some physiological indicators to test the performance of the model.

In this section, only 10 physiological variables including HR, GCS, NIDiasABP, NISysABP, PaCO2, PaO2, pH, Temp, Urine, and WBC are selected over repeated trials for JITL-ELM modeling. The ROC curve in Fig. 4 as well as the quantitative results in Table 6 confirmed that a small number of key variables can also achieve a good effect of mortality prediction, and the results are also better than other methods, especially than SAPS scoring system commonly used in the hospital currently, which illustrates a good performance and feasibility of simplified JITL-ELM.

By contrast, a brief introduction will be conducted about the SAPS-I. In terms of SAPS-I system, the worst physiological values in the first 24 h after patients entered ICU are collected, where the missing items are regarded as normal. Finally, 14 physiological variables are selected and scores are remarked for them, respectively. The higher the final score, the worse the condition and prognosis. However, the scoring system comes into being against the European, which may be not so suitable for Chinese patients.

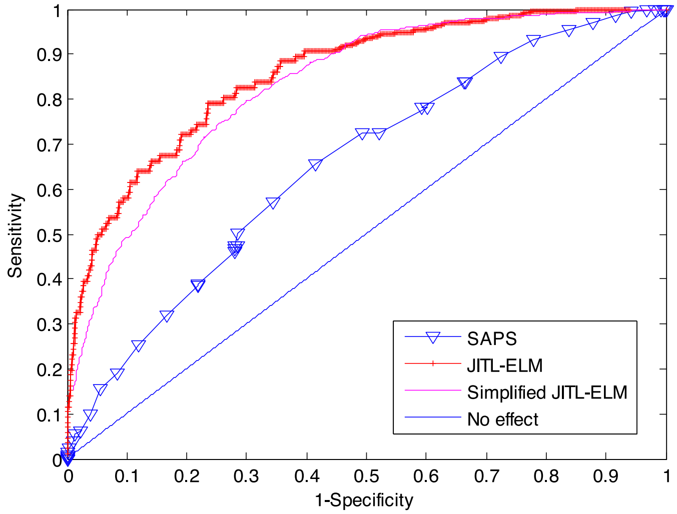


Fig. 4. ROC curves for JITL-ELM and simplified JITL-ELM.

As mentioned above, SAPS-I system needs to collect 14 variables to gain a relatively accurate results, while the simplified JITL-ELM only request 10 items. According to the SAPS-I score provided by the database, as shown in Fig. 4, the performance of the simplified JITL-ELM algorithm is still more competitive.

Additionally, the simplified JITL-ELM has a similar AUC indicator to ELM method in Table 6, but the simplified JITL-ELM uses less physiological variables, which makes it a more competitive approach than ELM.

What is worth mentioning, as the databases of appropriate patient information increase in size and complexity, the performance will be significantly improved since more useful information will be collected in similar dataset, which has a potential value to in clinical decision-making.

5. Conclusion

In this study, a novel combination, referred to JITL-ELM, is introduced and applied to the mortality prediction for ICU patients. The algorithm offers a general framework, in which JITL can search for a better domain space for testing samples and while ELM provides a fast learning method to get better prediction results. For further optimizing, a two-step scheme is proposed in this study, in which the first stage is for clustering and while the second one is for prediction. Possessing a high clinical value, it can narrow the search scope and improve the retrieval speed for JITL-ELM. Compared with the scoring system commonly used in hospitals currently, it promotes the classification accuracy significantly, especially better than the SAPS-I scoring system. Finally, the study tries

Please cite this article as: Y. Ding et al., Mortality prediction for ICU patients combining just-in-time learning and extreme learning machine, Neurocomputing (2017), https://doi.org/10.1016/j.neucom.2017.10.044

to establish a more practical model with less physiological variables, which is also performs much better than the SAPS-I system. Through tested on dataset collected from PhysioNet, the proposed algorithm has better performance compared with the traditional global modeling methods.

In summary, JITL-ELM can monitor individual patients with high adaptability and specificity. It has a potential application value for early warning systems in the future, which is also in step with the development trend of personalized medicine.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 61374099 and Research Fund for the Taishan Scholar Project of Shandong Province of China.

Competing interests

No competing financial interests exist.

References

[1] W.A. Knaus, E.A. Draper, D.P. Wagner, J.E. Zimmerman, APACHE II: a severity of disease classification system, Crit. Care Med. 14 (8) (1986) 754–755.
[2] G.J. Le, S. Lemeshow, F. Saulnier, A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study, J. Am. Med. Assoc. 270 (24) (1993) 2957–2963.
[3] S. Lemeshow, D. Teres, J. Klar, J.S. Avrunin, S.H. Gehlbach, J. Rapoport, Mortality probability models (MPM II) based on an international cohort of intensive care unit patients, JAMA: J. Am. Med. Assoc. 270 (20) (1993) 2478–2486.
[4] A.J. Hussain, P. Fergus, H. Al-Askar, D. Al-Jumeily, F. Jager, Dynamic neural network architecture inspired by the immune algorithm to predict preterm deliveries in pregnant women, Neurocomputing 151 (3) (2015) 963–974.
[5] K.J. Kim, S.B. Cho, Prediction of colon cancer using an evolutionary neural network, Neurocomputing 61 (1) (2004) 361–379.
[6] P. Chen, L. Yuan, Y. He, S. Luo, An improved SVM classifier based on double chains quantum genetic algorithm and its application in analogue circuit diagnosis, Neurocomputing 211 (2016) 202–211.
[7] A.T. Azar, S.A. El-Said, Performance analysis of support vector machines classifiers in breast cancer mammography recognition, Neural Comput. Appl. 24 (5) (2014) 1163–1177.
[8] A.T. Azar, S.M. El-Metwally, Decision tree classifiers for automated medical diagnosis, Neural Comput. Appl. 23 (7) (2013) 2387–2403.
[9] O.P. Ryynänen, E.J. Soini, A. Lindqvist, M Kilpeläinen, T. Laitinen, Bayesian predictors of very poor health related quality of life and mortality in patients with COPD, BMC Med. Inform. Decis. Mak. 13 (1) (2013) 1–10.
[10] Z. Cui, Y. Wang, X. Gao, J. Li, Y. Zheng, Multispectral image classification based on improved weighted MRF Bayesian, Neurocomputing 212 (2016) 75–87.
[11] M. Last, O. Tosas, T.G. Cassarino, Z. Kozlakidis, J. Edgeworth, Evolving classification of intensive care patients from event data, Artif. Intell. Med. 69 (2016) 22–32.
[12] J.G. Klann, P. Szolovits, S.M. Downs, G. Schadow, Decision support from local data: creating adaptive order menus from past clinician behavior, J. Biomed. Inform. 48 (3) (2014) 84–93.
[13] Z. Ying, P. Szolovits, Patient-specific learning in real time for adaptive monitoring in critical care, J. Biomed. Inform. 41 (3) (2008) 452–460.
[14] C.G. Enright, M.G. Madden, Modelling and Monitoring the Individual Patient in Real Time, Springer International Publishing, 2015.
[15] N. Kasabov, Y. Hu, Integrated optimisation method for personalised modelling and case studies for medical decision support, Int. J. Funct. Inform. Person. Med. 3 (3) (2010) 236–256.
[16] X. Li, Y. Wang, Adaptive online monitoring for ICU patients by combining just-in-time learning and principal component analysis, J. Clin. Monit. Comput. 30 (6) (2015) 1–14.
[17] Y. Ding, X. Li, Y. Wang, Mortality prediction for ICU patients using just-in-time learning and extreme learning machine, in: Proceedings of World Congress on Intelligent Control and Automation, 2016, pp. 939–944.
[18] G.B. Huang, Q.Y. Zhu, C.K. Siew, Extreme learning machine: a new learning scheme of feedforward neural networks, in: Proceedings of International Joint Conference on Neural Networks, 2, 2004, pp. 985–990.
[19] G.B. Huang, H. Zhou, X. Ding, R. Zhang, Extreme learning machine for regression and multiclass classification, IEEE Trans. Syst. Man Cybern. B: Cybern. 42 (42) (2012) 513–529.
[20] A.J. Mayne, Generalized Inverse of Matrices and its Applications, John Wiley&Sons, Inc., 1972.
[21] A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics 42 (1) (1970) 80–86.
[22] G.B. Huang, L. Chen, Convex incremental extreme learning machine, Neurocomputing 70 (16–18) (2007) 3056–3062.
[23] C. Cheng, M.S. Chiu, A new data-based methodology for nonlinear process modeling, Chem. Eng. Sci. 59 (13) (2004) 2801–2810.
[24] K. Fujiwara, M. Kano, S. Hasebe, Development of correlation-based clustering method and its application to software sensing, Chemom. Intell. Lab. Syst. 101 (2) (2010) 130–138.
[25] K. Chen, J. Ji, H. Wang, Y. Liu, Z. Song, Adaptive local kernel-based learning for soft sensor modeling of nonlinear processes, Chem. Eng. Res. Des. 89 (10) (2011) 2117–2124.
[26] Y. Liu, Z. Gao, P. Li, H. Wang, Just-in-time kernel learning with adaptive parameter selection for soft sensor modeling of batch processes, Ind. Eng. Chem. Res. 51 (11) (2012) 4313–4327.
[27] A. El-Hamdouchi, P. Willett, Hierarchic document custering using Ward's method, in: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa, Italy, September 1986, 1986, pp. 149–156.
[28] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (8) (2006) 861–874.
[29] M. Saeed, M. Villarroel, A.T. Reisner, G. Clifford, L.W. Lehman, G. Moody, et al., Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database, Crit. Care Med. 39 (5) (2011) 952–960.
[30] V.I. Pagurova, On Chauvenet's test for finding several outliers, Theory Probab. Appl. 30 (3) (1986) 558–561.
[31] Z. Lou, J. Tuo, Y. Wang, Two-step principal component analysis for dynamic processes, in: Proceedings of International Symposium on Advanced Control of Industrial Processes, 2017, pp. 73–77.

Yangyang Ding was born in 1991 in China. She received her Bachelor degree from Beijing University of Chemical Technology in 2015, majoring in Automation. She is currently working toward the Master degree at the same university. Her research interests include artificial neural networks, machine learning, and their biomedical applications.

Youqing Wang received his B.S. degree from Shandong University, Jinan, Shandong, China, in 2003, and his Ph.D. degree in control science and engineering from Tsinghua University, Beijing, China, in 2008. He worked as a Research Assistant in the Department of Chemical Engineering, Hong Kong University of Science and Technology, from February 2006 to August 2007. From February 2008 to February 2010, he worked as a senior investigator in the Department of Chemical Engineering, University of California, Santa Barbara, USA. From August 2015 to November 2015, he was a visiting professor in Department of Chemical and Materials Engineering, University of Alberta, Canada. Currently, he is a professor in Shandong University of Science and Technology and also Beijing University of Chemical Technology. His research interests include fault-tolerant control, state monitoring, modeling and control of biomedical processes (e.g. artificial pancreas system), and iterative learning control. He is an Associate Editor of *Multidimensional Systems and Signal Processing* and *Canadian Journal of Chemical Engineering*. He holds membership of two IFAC Technical Committees (TC6.1 and TC8.2). He is a recipient of several research awards (including Journal of Process Control Survey Paper Prize and ADCHEM2015 Young Author Prize).

Donghua Zhou received the B.Eng., M.Sci., and Ph.D. degrees in electrical engineering from Shanghai Jiaotong University, China, in 1985, 1988, and 1990, respectively. He was an Alexander von Humboldt research fellow with the University of Duisburg, Germany from 1995 to 1996, and a visiting scholar with Yale university, USA from 2001 to 2002. He joined Tsinghua University in 1996, and was promoted as a Full Professor in 1997, he was the head of the department of automation, Tsinghua university, during 2008 and 2015. He is now the Vice President, Shandong University of Science and Technology. He has authored and coauthored over 160 peer-reviewed international journal papers and 6 monographs in the areas of process identification, fault diagnosis, fault-tolerant control, reliability prediction, and optimal maintenance. Dr. Zhou is a member of the IFAC TC on SAFEPROCESS, a senior member of IEEE, an associate editor of the Journal of Process Control, the vice Chairman of Chinese Association of Automation (CAA), the Chairman of the national high education steering committee on automation, the TC Chair of the SAFEPROCESS committee, CAA. He was also the NOC Chair of the 6th IFAC Symposium on SAFEPROCESS 2006.

Please cite this article as: Y. Ding et al., Mortality prediction for ICU patients combining just-in-time learning and extreme learning machine, Neurocomputing (2017), https://doi.org/10.1016/j.neucom.2017.10.044