

Mortality Prediction of ICU Patient Based on Imbalanced Data Classification Model



Xuedong Gao, Hailan Chen, and Yifan Guo

Abstract The mortality prediction of Intensive Care Unit (ICU) patient has been an active topic in the medical filed. In recent years, data mining methods have been applied in this domain and have achieved good results. In this paper, we propose an imbalanced data classification algorithm to predict the mortality of ICU patient. Firstly, we introduce clustering algorithm to convert numeric data into categorical data. Then, imbalanced data processing method is used to obtain a balanced data. Finally, we apply some traditional classification algorithms to conduct experiment on a medical data set, and the experimental results show that the classification accuracy of balanced data set is more efficient than the imbalanced data set considering the minority death patients.

Keywords ICU · Mortality prediction · Imbalanced data · Classification

1 Introduction

Intensive Care Unit (ICU) is established to provide high quality medical treatment for critically ill patients. The patient mortality rate is one of the important factors to measure the ICU level. And effectively predicting the risk of death in ICU patients

This work is supported by national science fund of China (No. 71272161).

X. Gao · H. Chen (✉)
Donlinks School of Economics and Management, University of Science and Technology
Beijing, Beijing, China
e-mail: chl_hld@163.com

X. Gao
e-mail: gaoxuedong@manage.ustb.edu.cn

Y. Guo
School of Management, China University of Mining and Technology, Beijing, China
e-mail: guo_onlyIdol@163.com

can reasonably allocate rescue resources and reduce the cost of first aid. The traditional prediction methods are medical scoring systems such as APACHE [1], SAPS [2], SOFA [3], etc., which are difficult to meet clinical practice and not effectively deal with case-mix [4] problems. Therefore, how to accurately predict the mortality of ICU patient according to the early admission records, and take appropriate measures to improve the patient survival rate has become a research hot spot and difficulty in the medical field.

In recent years, with the improvement of medical informatization, most medical information has been preserved in digital form, providing a reliable and rich source of data for developing new mortality prediction methods. On the other hand, the rapid development of data mining and machine learning technology provides a theoretical basis for the development of new mortality prediction methods. Many researchers [5, 6] introduce classification algorithms into medical domain and have good results. Lin et al. [7] applied SVM (Support Vector Machine) to predict in-hospital mortality risk of patients with acute kidney injury in ICU. Xie et al. [8] used random forest model to predict in-hospital mortality risk of ICU patients. Zhang et al. [9] applied SVM, ANN and C4.5 to verify the predictive performance based on a selected time interval after learning classification knowledge for the same data set. Davoodia et al. [10] proposed the deep rule-based fuzzy classifier to predict mortality of ICU patient.

Among these studies, most of them assume the data set to be well-balanced and focus on achieving overall classification accuracy, which may lead to classification errors in minority class samples. However, most medical data sets are usually imbalanced, with more cases of survival and fewer deaths. And we are more concerned with the prediction of death. So how to deal with imbalanced data to improve the classification performance is essential. There are two main methods. The first is to improve the data set by increasing the number of minority class samples by over-sampling and decreasing the number of majority class samples by under-sampling [11]. The other method is to improve the algorithm itself. There have been many studies [12] to improve the classification accuracy of minority class samples by improving the algorithm. But these methods are not universally suitable for all data sets. It is easier and more efficient to converting an imbalanced data set into a balanced data set. Wu et al. [11] proposed the cluster-based under-sampling method to establish a balanced data set in the data processing stage. In this paper, we apply this method to process the imbalanced data set.

The rest of this paper is organized as follows. In Sect. 2 we describe the data and summarize its properties. In Sect. 3 we propose the mortality prediction model of ICU Patient. In Sect. 4 we conduct experiment and analyze the experimental results. Finally in Sect. 5 we give the conclusion.

2 Data Description

In this section, we describe the experimental data in this paper including data sources and data properties.

2.1 Data Sources

In this paper, the dataset we used is downloaded from PhysioNet website [13], which is the PhysioNet/Computing in Cardiology Challenge provided by of the National Institutes of Health in 2012. It contains 4000 cases from four different ICUs (Coronary Care, Cardiac Surgery, Recovery Units, Medical and Surgical); all cases collected at most 36 physiological variables; all hospitalized for more than 48 h and all over 16 years of age. There were a total of 554 deaths and 3346 surviving cases, and the missing rate of the different physiological variables ranged from 1.6 to 95.3%. Among all variables, we select 16 physiological variables as shown in Table 1, which the missing data rate is less than 20%.

The data for each ICU patient is generally composed of three components, the basic information at admission, the sequence of physiological variables after admission, and the final survival or death status. A typical ICU data set is shown in Table 2.

Table 1 Experimental physiological variables

Abbreviation	Full name	Unit
BUN	Blood urea nitrogen	mg/dL
Creatinine	Creatinine	mg/dL
GCS	Glasgow coma scale	–
Glucose	Glucose	mg/dL
HCO ₃	Serum bicarbonate	mmol/L
HCT	Red blood cell specific volume	%
HR	Heart rate	bpm
K	Kalium	mEq/L
Mg	Serum magnesium	mmol/L
Na	Sodium	mEq/L
NIDiasABP	Non-invasive diastolic arterial blood pressure	mmHg
NISysABP	Non-invasive systolic arterial blood pressure	mmHg
Platelets	Blood platelet count	cells/nL
Temp	Human body temperature	°C
Urine	Urine	mL
WBC	White blood cell	cells/nL

Table 2 Example of ICU data set

ID _n (demograph <vector>, physiological, outcome)			
HR	NIDiasABP	...	Temp
(00:00, 73)	(00:00, 88)	...	(00:00, 36.3)
(01:09, 67)	(00:09, 79)	...	(00:09, 35.8)
(01:24, 50)	(01:09, 80)	...	(05:09, 36.6)
(01:39, 63)	(01:39, 73)	...	(09:09, 36.8)
(01:54, 64)	(01:54, 76)	...	(13:09, 37.1)
...

2.2 Properties of the Data

After preliminary exploration of the data, we summarize the properties as follows: high dimensionality, large amount of missing data, imbalance data, and different time intervals.

- High dimensionality: The ICU data set is highly dimensional in two respects. One performance is high in attributes. It has 36 physiological variables, such as temperature, blood pressure, and heart rate etc. Another one performance is high in time. For a physiological variable of one patient, there are multiple observations within 48 h. It leads to the high dimensionality of time.
- Large amount of missing data: The data set contains a large amount of missing data, and the missing rate ranged from 1.6 to 95.3% for different physiological variables. When the missing rate of the data set is not high, we can use the missing data filling methods to fill the data; while when the missing rate is high, the commonly missing data filling methods have no effect.
- Imbalance data: The data set contains more samples from one class (3446 survival cases) while the other is much smaller (only 554 dead cases). It is severely imbalanced. The phenomenon may lead to classification errors in minority class samples. Therefore, the classifier may behavior too bad to get satisfactory results. It is necessary to handle the imbalanced problem.
- Different time intervals: As the sensors and devices may break down, or medical staff may incorrectly operate these medical equipment, the variables of each record were sampled unevenly. For one variable of one patient, the time intervals of each two records may be different. It also can be regarded as a reflection of missing data.

Therefore, there is an urgent need to solve the above issues and develop an effective method to improve the mortality prediction accuracy of ICU patient.

3 Mortality Prediction Model of ICU Patient

In this section, we first propose the framework of ICU mortality prediction model. Then according to this framework, we introduce the method of data preprocessing, imbalanced data converting and classification. At last, we present evaluate indexes of prediction results.

3.1 The Framework of ICU Mortality Prediction Model

How to predict the mortality of ICU patient is meet with many challenges and difficulties. So, in this subsection we propose the framework of ICU mortality prediction model as shown in Fig. 1.

According to the proposed framework, this prediction model mainly includes four steps: data preprocessing, imbalanced data processing, classification, and evaluation.

3.2 Data Preprocessing

Since the ICU data have large amount of missing data and high dimensions, in this subsection, we preprocess the data to solve these challenges.

In order to avoid the instability caused by filling missing data, we select 16 physiological variables as shown in Table 1, which the missing data rate is less than 20%. These indexes are also considered to be closely related to death in medical. In this way, we reduce the attribute dimensions.

Fig. 1 The framework of ICU mortality prediction model

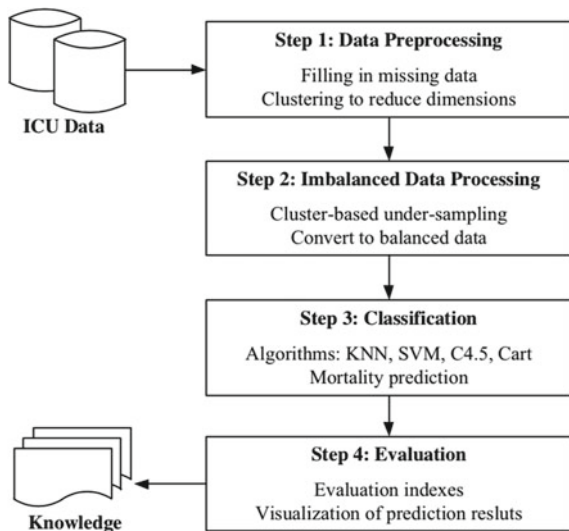


Table 3 Two-dimension information matrix

ID	HR	NIDiasABP	...	Temp	Outcome
132597	4	10	...	7	0
132598	1	1	...	9	1
132599	7	9	...	9	0
132601	8	5	...	2	0
132602	9	1	...	2	1
132605	9	8	...	1	1
132610	9	10	...	1	0

In Table 2, for a physiological variable of one patient, there are multiple observations within 48 h. It leads to the high dimensionality of time. So, we calculate the mean value of multiple observations for each attribute within 48 h to represent the attribute. And the missing value of each attribute is replaced by the mean of the attribute.

After the above operation, we get a matrix in which each patient has only one value per attribute. However, the variation range of each attribute is very different, we apply clustering algorithm to convert numeric data into categorical data. K-means clustering algorithm is used to do this operation, and we set $k = 10$. Completing clustering, the ICU data is converted into a matrix of two-dimension information (Table 3).

3.3 Imbalanced Data Processing Method Based on Cluster Under-Sampling

In this subsection, we introduce an imbalanced data processing method based on cluster under-sampling to improve the prediction accuracy.

The algorithm steps are as follows.

Input: Imbalanced data set.

Step 1: Divide the original data set into one class of minority data samples and the other class of majority data samples. And directly put the class of minority data samples is into a new data set.

Step 2: Calculate the number of samples in the minority class, which is set to the number of majority clusters k .

Step 3: K-means algorithm is used to cluster the majority data samples of another class. And Euclidean distance is selected as the distance measurement.

Step 4: After clustering, the center points of k clusters are put into the new data set. At the moment, the data samples of two classes are equal, and the new data set is balanced.

Output: Balanced data set.

As shown in Figs. 2 and 3, Fig. 2 is the data distribution before clustering, and Fig. 3 is the data distribution after clustering. Although the amount of data after

clustering is reduced, but the spatial distribution of the data set remains unchanged, and the distribution characteristics of the data set are preserved.

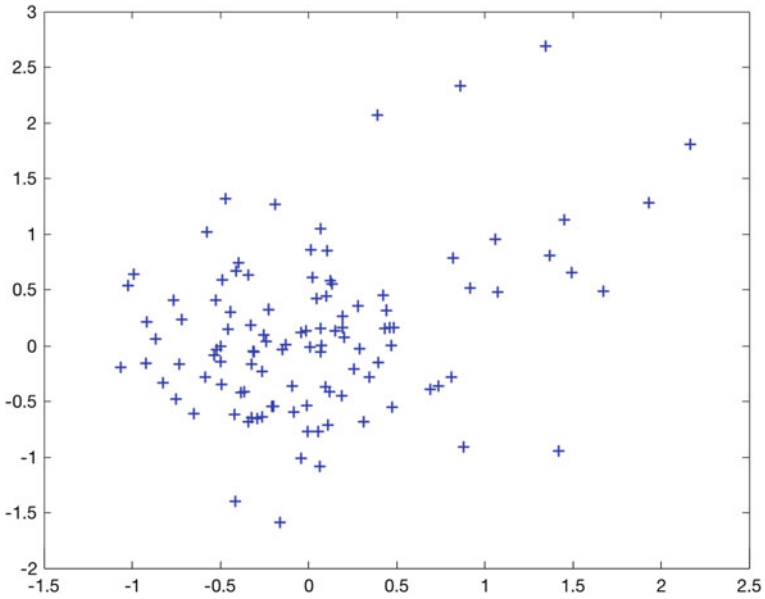


Fig. 2 Data distribution before clustering

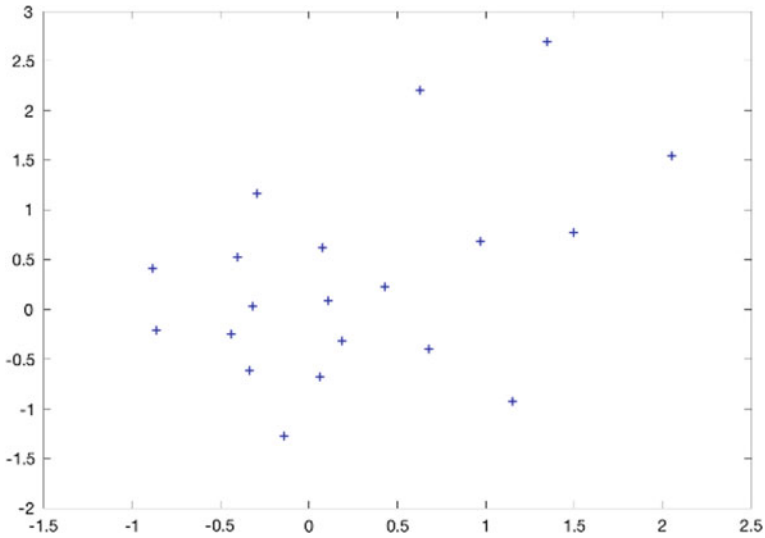


Fig. 3 Data distribution after clustering

3.4 Classification Algorithms

Classification is an important data mining technique. Classification is usually used for prediction. The classification algorithms mainly include decision tree, statistical method, machine learning method, and neural network method [14]. In this paper, the classification algorithms we used are K-Nearest Neighbor (KNN) algorithm, Support Vector Machine (SVM) algorithm, C4.5 algorithm, and Classification and Regression Tree (CART) algorithm. They are traditional classification algorithms.

KNN algorithm is one of the best classification algorithms under the vector space model (VSM) [15]. The basic idea is to calculate the similarity between the sample to be classified and the training sample under the VSM representation. Finding the k nearest neighbors that are most similar to the sample to be classified, and determining the attribution of the samples to be classified according to the classes of the k neighbors.

As a classifier, SVM has proved itself being a powerful tool for classification and regression [7]. The goal of the SVM is to create an optimal classification boundary (maximum interval hyperplane) in high-dimensional space to distinguish between different classes of samples. The support vector refers to the point in each class that is closest to the largest interval hyperplane.

C4.5 is currently the most influential decision tree classification algorithm. The C4.5 algorithm uses information gain rate instead of information gain to select node attributes, supports discrete attributes and continuous attributes, and is a decision tree learning algorithm based on information entropy [16].

The CART algorithm uses the minimum Gini value as the demarcation point selection criterion, and is a binary tree. It has the advantages of small computation, fast speed and easy interpretation [17]. The algorithm works well for a large amount of real data in a relatively short time.

3.5 Evaluation Indexes

Currently, there are many evaluation indexes to measure the classification problem. In this work, we use confusion matrix to evaluate ICU mortality prediction. This matrix is composed of true positive (TP), true negative (TN), false positive (FP) and false negative (FN).

For the ICU data set, we define the death class (positive class) sample as P, and the survivor class (negative class) sample as N. The specific confusion matrix is shown in Table 4. The calculation indexes based on the confusion matrix are shown

Table 4 Definition of confusion matrix

Outcome		Observed	
		Death	Survivor
Predicted	Death	TP	FP
	Survivor	FN	TN

in Eqs. (1) to (7), which are TPR (true rate) or Recall, TNR (true negative rate) or Sp (specificity), FPR (false positive rate), FNR (false negative rate), Precision, Accuracy, F-measure.

$$TPR(Recall) = TP/(TP + FN) \quad (1)$$

$$TNR(Sp) = TN/(TN + FP) \quad (2)$$

$$FPR = FP/(TN + FP) \quad (3)$$

$$FNR = FN/(TP + FN) \quad (4)$$

$$Precision = TP/(TP + FP) \quad (5)$$

$$Accuracy = (TN + TP)/(TN + TP + FP + FN) \quad (6)$$

$$F\text{-measure} = (2 * Precision * Recall)/(Precision + Recall) \quad (7)$$

Except for the above evaluation indexes, in this paper, we introduce min (Recall, Precision) [9] to measure the prediction results. Because it better reflects the classification accuracy of the class with few samples.

4 Experiments

In this section, we use MySQL to extract the data, data preprocessing and prediction model building is written by MATLAB.

In this experiment, we divide the ICU data set into training set and test set, and the ratio is 3:1. Then we calculate the performance indexes of each classification algorithm. In order to verify the effectiveness of the proposed method, we use the ICU data set without doing imbalanced data processing as comparison experiments to conduct classification. For imbalanced data set, there are two classes (554 dead cases and 3446 survival cases), and its corresponding ratio is 1:6.22. So it has 3000 cases in training set and 1000 cases in test set. For balanced data set, the number of two classes is equal (554 dead cases and 554 survival cases). So it has 831 cases in training set and 277 cases in test set. And the results is shown in Table 5.

Table 5 Performance comparison of four classification algorithms on imbalanced data set and balanced data set

Classification algorithms		Evaluation indexes				
		Recall/%	Precision/%	min (Recall, Precision)/%	F-measure/%	Accuracy/%
KNN	Imbalanced data set	10.48	18.33	10.48	13.33	85.70
	Balanced data set	42.45	54.63	42.45	47.77	53.43
SVM	Imbalanced data set	11.39	11.76	11.39	11.57	89.50
	Balanced data set	83.45	78.57	78.57	80.94	79.42
C4.5	Imbalanced data set	35.24	24.50	24.5	28.91	81.80
	Balanced data set	72.66	74.81	72.66	73.72	74.01
CART	Imbalanced data set	24.39	23.26	23.26	23.81	85.35
	Balanced data set	81.13	78.18	78.18	79.63	78.74

In Table 5, five evaluation indexes (Recall, Precision, min (Recall, Precision), F-measure, Accuracy) are used to present the performance of four classification algorithms (KNN, SVM, C4.5, and CART) on imbalanced data set and balanced data set. It is obvious to see that balanced data set achieves better performance on four indexes: Recall, Precision, min (Recall, Precision), F-measure. We also see that the performance of imbalanced data set is not good on these four indexes. However, its accuracy is high. This phenomenon indicates that these classification algorithms are greatly affected by the imbalance of data, and tend to divide objects into classes with majority samples, ignoring the class with minority samples.

Although the accuracy of balanced data set is lower than using imbalanced data set, the difference of them are not obvious. When we classify the imbalanced data, we pay more attention to the accuracy of the class with majority samples, which is consistent with our real life.

The above experimental results demonstrate that the quality of the proposed framework of mortality prediction model is effective.

5 Conclusion

In this paper, we summarize the properties ICU data: high dimensionality, large amount of missing data, imbalance data, and different time intervals. In order to solve these challenges, we propose the framework of ICU mortality prediction model. In this framework, we first preprocess the data. We select some variables from many attributes, meanwhile we fill the missing data. And K-means clustering is applied to convert numeric data into categorical data. After completing the data preprocessing, we introduce an imbalanced data processing method based on cluster under-sampling to improve the prediction accuracy. Using this method, we can convert imbalanced data into balanced data. And it also reduces the quantity of data set. Then we present the tradition classification algorithms. Furthermore, we give the evaluation indexes of classification accuracy. Finally, experiments are carried out to evaluate the performance of the proposed framework of ICU mortality prediction model. And the experimental results verify that this model works effectively.

References

1. Zimmerman, J. E., Kramer, A. A., & McNair, D. S. (2006). Acute physiology and chronic health evaluation IV: Hospital mortality assessment for today's critically ill patients. *Critical Care Medicine*, 34, 1297–1310.
2. Gall, J. R. L., Lemeshow, S., & Saulnier, F. (1993). A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *Journal of the American Medical Association*, 270, 2957–2963.
3. Vincent, J. L., Moreno, R., & Takala, J. (1996). The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22, 707–710.
4. Strand, K., & Flaatten, H. (2008). Severity scoring in the ICU: A review. *Acta Anaesthesiologica Scandinavica*, 52, 467–478.
5. Liu, S. J., Chen, X. X., Fang, L., Li, J. X., Yang, T., Zhan, Q., et al. (2018). Mortality prediction based on imbalanced high-dimensional ICU big data. *Computers in Industry*, 98, 218–225.
6. Xu, J., Zhang, Y., Zhang, P., Mahmood, A., Li, Y., & Khatoun, S. (2017). Data Mining on ICU Mortality prediction using early temporal data: A survey. *International Journal of Information Technology & Decision Making*, 16, 117–159.
7. Lin, K., Xie, J. Q., Hu, Y. H., & Kong, G. L. (2018). Application of support vector machine in predicting in-hospital mortality risk of patients with acute kidney injury in ICU. *Journal of Peking University (Health Sciences)*, 50, 239–244.
8. Xie, J. Q., Lin, K., Li, C. X., & Kong, G. L. (2017). Application of random forest model in the prediction of in-hospital mortality risk in ICU Patients. *China Digital Medicine*, 12, 81–84.
9. Xu, J., Zhang, Y., Zhou, D., Li, D., & Li, Y. (2014). Uncertain multi-granulation time series modeling based on granular computing and the clustering practice. *Journal of Nanjing University (Natural Sciences)*, 50, 86–94.
10. Davoodi, R., & Moradi, M. H. (2018). Mortality prediction in intensive care units (ICUs) using a deep rule-based fuzzy classifier. *Journal of Biomedical Informatics*, 79, 48–59.

11. Wu, S., Liu, L., & Lu, D. (2017). An Imbalanced data ensemble classification based on cluster-based under-sampling algorithm. *Chinese Journal of Engineering*, 39, 1244–1253.
12. Tao, X. M., Hao, S. Y., Zhang, D. X., & Xu, P. (2013). Overview of classification algorithms for unbalanced data. *Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition)*, 25, 101–111.
13. Predicting Mortality of ICU Patients: The PhysioNet/Computing in Cardiology Challenge 2012. National Institutes of Health NIH, 2012. <http://physionet.org/challenge/2012/>. Accessed 27 Aug 2012.
14. Luo, K., Lin, M. G., & Xi, D. M. (2005). Review of classification algorithms in data mining. *Computer Engineering*, 31(1), 3–11.
15. Geng, L. J., & Li, X. Y. (2014). Improvements of KNN algorithm for big data classification. *Application Research of Computers*, 31, 1342–1344.
16. Cherfi, A., Noura, K., & Ferchichi, A. (2018). Very fast C4.5 decision tree algorithm. *Applied Artificial Intelligence*, 32, 119–137.
17. Xu, X. C., Saric, Z., & Kouhpanejade, A. (2014). Freeway incident frequency analysis based on cart method. *Promet - Traffic & Transportation*, 26(3), 191–199.