MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231081

Extracting Dynamic Information of Temporal Clinical Data to Predict the Outcome in Critically Ill Patients

Jing XIA[a], Yi REN[a], Zhenchuan ZHANG[a], Feng WANG[a], Yu TIAN[b], Tianshu ZHOU[a] and Jingsong LI[a,b,1]
[a] *Research Center for Healthcare Data Science, Zhejiang Laboratory, Hangzhou, China*
[b] *Engineering Research Center of EMR and Intelligent Expert System, Ministry of Education, College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, China*
ORCiD ID: Jing Xia https://orcid.org/0000-0002-5476-727X

Abstract. Outcome prediction is essential for the administration and treatment of critically ill patients. For those patients, clinical measurements are continuously monitored and the time-varying data contains rich information for assessing the patients' status. However, it is unclear how to capture the dynamic information effectively. In this work, multiple feature extraction methods, i.e. statistical feature classification methods and temporal modeling methods, such as recurrent neural network (RNN), were analyzed on a critical illness dataset with 18415 cases. The experimental results show when the dimension increases from 10 to 50, the RNN algorithm is gradually superior to the statistical feature classification methods with simple logic. The RNN model achieves the largest AUC value of 0.8463. Therefore, the temporal modeling methods are promising to capture temporal features which are predictive of the patients' outcome and can be extended in more clinical applications.

Keywords. Time-varying data, temporal features, outcome prediction, critically ill

1. Introduction

Critically ill patients are usually continuously monitored in intensive care units (ICU). The outcome prediction is important for those patients' management and treatment [1]. The temporal data obtained through continuous monitoring can provide valuable information for doctors to assess a patient's condition and track changes over time. However, analyzing this data can present challenges due to its complexity and volume. In recent decades, more and more data-driven risk prediction models are used to assist clinical disease assessment, such as the outcome prediction task [1,2].

To handle time-varying variables, there are generally two kinds of approaches. The first is statistical feature classification methods, which extract statistical features from sequence data manually and input to classification algorithms, such as logistic regression (LR) and random forests (RF). The APACHE II scoring system [3], for

[1] Corresponding Author: Jingsong Li, email: ljs@zju.edu.cn.

example, uses the worst value in the previous 24 hours for evaluating a patient's status. Previous studies extracted the last moment value, mean, standard deviation, minimum and maximum for continuous-valued variables [3-5]. Such methods discard part of the time-varying information and oversimplify the relationship between variables.

Another is temporal modelling methods, which capture dynamic information from the temporal data using model itself and build prediction models directly on the time series data. Some researchers applied recurrent neural network (RNN) to capture underlying temporal structures [6], such as long-short term memory (LSTM) and gated recurrent unit (GRU), while others adopted convolutional neural network (CNN) to capture local sequence features [7] or the fusion framework convolutional-RNN [8].

Presently, it is unclear which kind of approach is better in handling time-varying variables and under what conditions is a particular algorithm superior. In addition, it is doubtful whether the integration of existing temporal feature extraction technologies can promote the outcome prediction performance. Therefore, in this work, multiple temporal feature extraction and classification methods are compared and discussed.

2. Methods

2.1. Data and preprocessing

Medical Information Mart for Intensive Care III (MIMIC-III) database [9] is a large and publicly available database of ICU admissions at the Beth Israel Deaconess Medical Center, USA, from 2001 to 2012. We included patients with age >15 years and length of ICU stay ≥ 10 days. The outcome is 28-day mortality, i.e. whether or not a patient dies within 28 days after ICU admission. Totally 18415 ICU records were enrolled, consisting of 2162 positive cases and 16253 negative ones. And 50 important and commonly used clinical measurements[1] were extracted for 10 consecutive days after admission, resulting in a dataset of sequence length 10. Missing values in data were filled by linear interpolation method and all variables were normalized by $\frac{x-mean}{std}$.

2.2. Development of the outcome prediction models

Classic non-sequential algorithms LR/RF and neural network-based sequential algorithms are used for classification. And the dynamic information is captured statistically or by algorithm itself. Hence, the prediction methods were as follows.

2.2.1. Statistical feature classification methods

- LR_0/RF_0: the last moment values are input to LR/RF for classification.
- LR_1/RF_1: statistical features (mean, std, maximum, minimum, magnitude of the whole sequence, last moment value, change of the last two time points) are input to LR/RF for classification.

2.2.2. Temporal modelling methods

- CNN: a conv-1D layer (16 filters with kernel size=3) and then a fully connected (FC) layer.

- RNN: a GRU layer (hidden size=16) and then a FC layer.
- ConvRNN: a conv-1D layer (16 filters with kernel size=3), a GRU layer (hidden size=16), and then a FC layer.

2.2.3. Fusion methods

- stat_dsRNN: statistical features (same as LR_1) are concatenated with the output of the last timestep from GRU layer, then input to a FC layer.
- CNN_dsRNN: static features learned from CNN model and the output of the last timestep from the GRU layer are input to a FC layer.
- RNN_withCNN: dynamic features learned from CNN and the original time series are input to a GRU layer, followed by a FC layer.

For neural network-based methods, learning rate is 0.01 and weight decay is 0.005. The loss function is the cross entropy between the true label and the predicted score. Adam optimizer and early stop strategy is used for training.

2.3. Performance evaluation

To evaluate the performances of prediction models, all samples were randomly split into a training set (90% of samples) and a test set (remaining 10% samples). The experiment was repeated multiple times. The area under the receiver operating characteristic curve (AUC), sensitivity, specificity and accuracy were reported.

3. Results

3.1. Comparison of multiple feature extraction and classification methods

The performances of multiple feature extraction and classification methods are shown in Table 1. For the key metric AUC, the RNN algorithm achieves the largest AUC value of 0.846, while the LR_0 has the smallest AUROC value of 0.828. The other algorithms perform slightly worse than RNN in terms of AUROC. Besides, the RF-based prediction models are competitive with RNN in sensitivity. The ConvRNN algorithm has slightly higher specificity and accuracy than RNN.

Table 1. Performances of multiple feature extraction and classification methods.

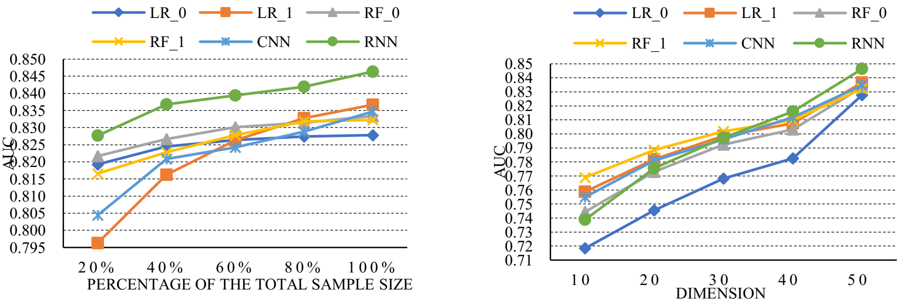
Type of method	**Algorithm**	**AUC (mean±std)**	**Sensitivity (mean±std)**	**Specificity (mean±std)**	**Accuracy (mean±std)**
Statistical feature classification methods	LR_0	0.828±0.016	0.748±0.024	0.759±0.013	0.758±0.011
	LR_1	0.837±0.018	0.718±0.049	0.793±0.020	0.784±0.014
	RF_0	0.832±0.011	0.767±0.028*	0.746±0.016	0.749±0.013
	RF_1	0.834±0.012	0.765±0.030*	0.754±0.020	0.755±0.016
Temporal modelling methods	CNN	0.835±0.017	0.726±0.037	0.781±0.013	0.774±0.012
	RNN	0.846±0.015*	0.758±0.028	0.770±0.015	0.768±0.013
	ConvRNN	0.835±0.013	0.718±0.034	0.794±0.014*	0.785±0.012*
Fusion methods	stat_dsRNN	0.839±0.016	0.750±0.036	0.768±0.017	0.766±0.014
	CNN_dsRNN	0.836±0.015	0.721±0.049	0.789±0.020	0.781±0.017
	RNN_withCNN	0.840±0.013	0.736±0.028	0.783±0.017	0.778±0.014

* marks the largest value of a metric.

3.2. Influence of the sample size and the dimension on predictive performances

The relationship between the sample size and the prediction performances are shown in Figure 1(a). As the sample size decreases, the AUC of LR_0 and RF-based models gradually decrease, while sequential models (CNN, RNN) exhibit a more rapid decline. The AUC of RNN is at least 0.01 higher than traditional models on the whole dataset, while that is only 0.005 higher on the dataset with a percentage of 20% samples. When the sample size ratio is 20%, the AUC of LR_1 drops to 0.795, probably because high dimension and small sample size lead to poor fitting of LR.

The relationship between the dimension and the prediction performances are summarized in Figure 1(b). For the dataset with dimension 10, most prediction models have similar AUC values, except LR_0. As the dimension of the dataset increases from 10 to 50, the AUC values of the traditional methods (LR_1, RF_0, RF_1) increase by 0.06-0.09, while that of the RNN shows a significantly larger increase of 0.11.



(a) (b)

Figure 1. AUC values of multiple models on datasets with different (a) samples sizes, (b) dimensions.

4. Discussion

In this study, multiple feature extraction strategies to handle the temporal data in critically ill patients were compared. The experimental results in Table 1 reveal that RNN is the best, probably because it captures useful global temporal features. The statistical methods (LR_1, RF_1) are competitive, nevertheless, it requires the guidance of prior knowledge to extract statistical features manually. And using the last moment value (LR_0, RF_0) does not perform well. Besides, CNN and ConvRNN do not show advantage, possibly because the local features here have little effect. Additionally, the fusion methods are not superior than RNN. Overall, RNN provides the most accurate outcome prediction for critically ill patients.

In Figure 1, with the sample size ratio reduced to 20%, the AUCs of the sequential models have a larger decline (RNN: 0.02, CNN: 0.03) than that of most classic methods (LR_0: 0.075, RF_0: 0.01, RF_1: 0.015), indicating the sample size has more impact on sequential methods than statistical feature classification methods. Moreover, as the dimension of the dataset rises, RNN is gradually superior to other algorithms with simple logic. Generally, RNN is more advantageous for outcome prediction in cases with more samples and a larger dimension.

5. Conclusions

For critically ill patients, time-varying data contains rich information for outcome prediction. Statistical methods and temporal modeling methods were compared on the MIMIC-III database. The experimental results demonstrate the temporal modeling method RNN has the best prediction performance. When the sample size or the dimension becomes small, statistical methods would be competitive. Generally, RNN is promising in capturing dynamic information from clinical temporal data.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62103382), the Zhejiang Provincial Natural Science Foundation of China (No. LQ21F030004), the China Postdoctoral Science Foundation (No. 2021M692959), the Key Research Project of Zhejiang Lab (No. 2022ND0AC01) and the Key Program of Science and Technology Development Program of Hangzhou, China (No. 2020ZDSJ0885).

References

[1] Shillan D, Sterne JA, Champneys A, Gibbison B. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. Crit Care. 2019 Dec;23:1-1, doi: 10.1186/s13054-019-2564-9.

[2] Goldstein BA, Pencina MJ. Developing Implementable Risk Prediction Models with Electronic Health Records Data. Wiley StatsRef: Statistics Reference Online. 2014 Apr;14:1-8, doi: 10.1002/9781118445112.stat08204.

[3] Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med. 1985 Oct;13(10):818-29.

[4] Hyland SL, Faltys M, Hüser M, Lyu X, Gumbsch T, Esteban C, Bock C, Horn M, Moor M, Rieck B, Zimmermann M. Early prediction of circulatory failure in the intensive care unit using machine learning. Nat Med. 2020 Mar;26(3):364-73, doi: 10.1038/s41591-020-0789-4.

[5] Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, Mottram A, Meyer C, Ravuri S, Protsyuk I, Connell A. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature. 2019 Aug;572(7767):116-9, doi: 10.1038/s41586-019-1390-1.

[6] Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc. 2017 Mar;24(2):361-70, doi: 10.1093/jamia/ocw112.

[7] Razavian N, Marcus J, Sontag D. Multi-task prediction of disease onsets from longitudinal laboratory tests. InMachine learning for healthcare conference 2016 Dec 10 (pp. 73-100). PMLR.

[8] Lin C, Zhang Y, Ivy J, Capan M, Arnold R, Huddleston JM, Chi M. Early diagnosis and prediction of sepsis shock by combining static and dynamic information using convolutional-LSTM. In2018 IEEE International Conference on Healthcare Informatics (ICHI) 2018 Jun 4 (pp. 219-228). IEEE, doi: 10.1109/ICHI.2018.00032.

[9] Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG. MIMIC-III, a freely accessible critical care database. Sci Data. 2016 May;3(1):1-9, doi: 10.1038/sdata.2016.35.

Endnotes

[1] The clinical measurements include SBP, HR, T, MAP, RR, GCS, A-aDO2, PaO2, FiO2, P/F, WBC, HCT, Hemoglobin, Plt, Albumin, RBC, RDW, HCO3-, Na+, K+, Chloride, Anion Gap, Glucose, Magnesium, Calcium, Phosphate, Alkaline Phosphatase, TBil, BUN, Cr, PH, Lactate, MCHC, MCH, MCV, INR, PT, PTT, Lymphocytes, Monocytes, Neutrophils, Basophils, Eosinophils, Base Excess, Calculated Total CO2, PCO2, Specific Gravity, ALT, AST, and PEEP.