Early Prediction of Sepsis from Clinical Data via Heterogeneous Event Aggregation

Luchen Liu[1*], Haoxian Wu[1*], Zichang Wang[1], Zequn Liu[1], Ming Zhang[1]

[1] Computer Science, Peking University, Beijing, China

Abstract

Sepsis is a life-threatening condition that seriously endangers millions of people over the world. Hopefully, with the widespread availability of electronic health records (EHR), predictive models that can effectively deal with clinical sequential data increase the possibility to predict sepsis and take timely preventive treatment. However, the timely prediction is challenging because patients' sequential data in EHR contains temporal interactions of multiple clinical events. And capturing temporal interactions in the long event sequence is hard for traditional LSTM. Rather than directly applying the LSTM model to the event sequences, our proposed model firstly aggregates heterogeneous temporal clinical events in a short period and then captures temporal interactions of the aggregated representations with LSTM. Our proposed Heterogeneous Event Aggregation can not only shorten the length of clinical event sequence but also help to retain temporal interactions of both categorical and numerical features of clinical events in the multiple heads of the aggregation representations. In the PhysioNet/Computing in Cardiology Challenge 2019, we have run the official scoring code with the following results of different metrics: AUROC (0.866), AUPRC (0.293), Accuracy (0.89), and Utility (0.402).

1. Introduction

Sepsis is a life-threatening condition that arises when the body's response to infection causes injury to its tissues and organs. And the early prediction of the sepsis onset is important for physicians to take timely preventive treatment. However, sepsis prediction is a difficult task, because complex risk factors range from a very young or elder age to a weakened immune system from conditions such as cancer, diabetes, major trauma, or burns. Hopefully, with the help of the widespread availability of electronic health records (EHR), predictive models that can effectively make use of clinical sequential data increase the accuracy of sepsis prediction performance.

The timely sepsis prediction is challenging because patients' sequential data in EHR contains temporal interactions of multiple clinical events[1, 2]. The interactions of multiple clinical events include event co-occurrence in a short period (e.g. two related symptoms occur together) and event temporal dependency at large time-scale (e.g. A vital signal abnormally arises several hours after certain drug injection). One possible solution is directly applying deep sequential models, such as LSTM[3], on the clinical event sequence. However, capturing temporal interactions in the long event sequence is hard for traditional LSTM for the length of clinical sequences that exceeds the modeling ability of LSTM.

Rather than directly applying the LSTM model to the event sequences, some works design hierarchical neural networks to model the long sequence[4]. For example, aggregating events in a short period into a vector helps to shorten the original long sequence[5]. However, the information of each kind of clinical events is mixed in the aggregation vector, so temporal interactions of these events are hard to capture.

To address these issues, our proposed model firstly aggregates heterogeneous temporal clinical events in a short period and then captures temporal interactions of the aggregated representations with LSTM. The Heterogeneous Event Aggregation module can not only shorten the length of clinical event sequence but also help to retain temporal interactions of both categorical and numerical features of clinical events in the multiple heads of the aggregation representations. The separated clinical information in different heads makes it easier to capture event temporal interactions in different aggregation vectors. Experiments on the PhysioNet/Computing in Cardiology Challenge 2019 show that our proposed model is effective and efficient compared to traditional methods.

2. Dataset and Preprocessing

2.1. Dataset

The EHR data provided publicly for this challenge is sourced from two separate ICU patients,containing 20000

* The two authors have equal contribution to this work.

and 20643 records respectively.Each record is made up of hourly clinical data for a specific patient.Each row represent a single hour's data with 40 column-variables and an additional label indicating whether the patient will get sepsis within 6 hours. There are 2932 sepsis patients totally.Firstly, sepsis and normal patients are divided into train and test set at the same ratio respectively. The larger 2 sets are arranged as a train set, and the remained sets are arranged as a test set. The training set is further divided into 5 groups by the same rule to construct cross validation sets.

2.2. Data preprocessing

In this competition, our goal is to early detecting sepsis within 6 hours for every single hour's monitoring data.As the data monitored posterior to the target hour is cannot be utilized for prediction, we sample data for each single hour's record. Supposing $patient_i$ has t_i rows, for every row_j of the record, we collect the data from row_{j-L+1} to row_j(zero filling if $j < L$) as the eventual data $D \in R^{L,40}$ for the j^{th} hour record of $patient_i$.

The 40 columns of the records contain 37 numerical variables and 3 two-categorical variables.Different columns have disparate sampled rates.At the preprocessing stage, we relabel the 3 two-categorical variables to 1-7(with one additional for NaN).Finally we reorder the categorical variables to the last columns. As for numerical variables, in order to make the deep model converge easily, we normalize each column of the numerical variables within mean value μ_i and standard deviation σ_i over all records for numerical $column_i$.For the i^{th} column of the t^{th} row of record D, we get the normalization data $N_{t,i} = (D_{t,i} - \mu_i)/\sigma_i$. Finally, we get the processed data $NinR^{n,L,40}$(n is total numbers of rows, L is sampled length of time-window)

3. Proposed Model

In this section, Attention-based sequential representation model is proposed to early detect sepsis as explained above. Heterogeneous Event Aggregation (HEA) module is designed to effectively capture the interaction information between the heterogeneous clinical events. After the dense representation of the sequential data is extracted, it will be passed on to a one-layer bidirectional LSTM to get early sepsis detect result.

The motivations of our proposed Heterogeneous Event Aggregation module are listed following: (1) Modeling interaction of both categorical and numerical heterogeneous features through event representation. (2) Grouping events into multiple heads to temporal interaction of different events. (3) Shortening the length of clinical event sequence

The proposed architecture is shown in figure 1.Given a

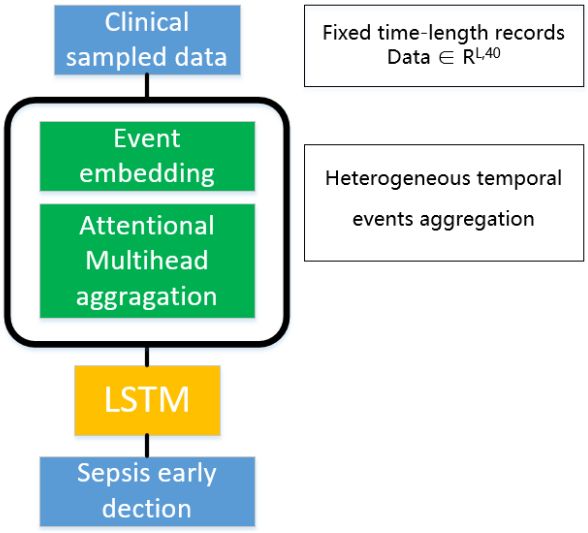

Figure 1. The architecture of our proposed model

sequential clinical record$(X_1, X_2, X_3...X_L), X \in R^{L,40}$, our objective is to generate early prediction of sepsis for the last time step x_L.To capture the useful information of complex interaction among clinical events, Heterogeneous Event Aggregation module is proposed.The module is composed of two steps Events Embedding and Attentional Multihead Aggregation.

3.1. Events Embedding

: Given the sequential clinical data, the first step of our model is to generate the embedding that can be used to capture the interaction representation among the heterogeneous clinical events[6]. For each time step $X_t \in R^{40}$(the first 37 columns are numerical variables,and the last 3 are categorical variables). Random initialized numerical event vector book $W_{ne} \in R^{37,d}$, categorical event lookup table $W_{ce} \in R^{7,d}$ and value vector table $W_v \in R^{37,d}$ are generated. The embedding for X_t is then generated as:

$$E_t = Mask(Concat(En_t, Ec_t)) \quad (1)$$
$$En_t = W_{ne} + X_t[:37] \cdot W_v \quad (2)$$
$$Ec_t = Embedding_lookup(X_t[37:], W_{ce}) \quad (3)$$
$$K_t = Concat(W_{ne}, Ec_t) \quad (4)$$

d is the dimensional numbers of embedding $E_t \in R^{40,d}$ is eventual event embedding, $En_t \in R^{37,d}$ is nurmerical embedding, $Ec_t \in R^{3,d}$ is categorical embedding, $K_t \in R^{40,d}$ is Key vector of both numerical and categorical events.

3.2. Attentional Events Aggregation

:

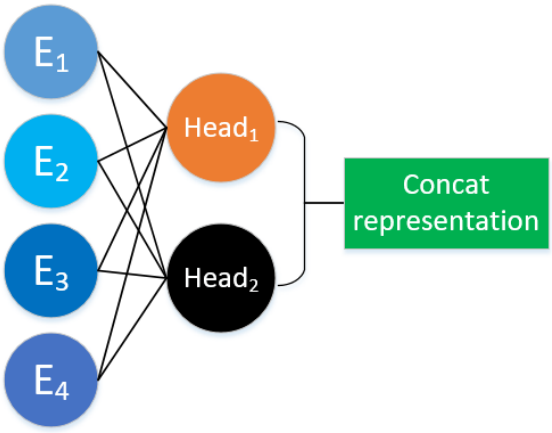

Figure 2. An example of two-heads events aggregation

Different variables have disparate sampled ratio, with diverse heterogeneous temporal events, there is abundant information in the long sequential record. It is difficult to extract the information through the long sequential data for the two reasons. (1). Within the long sequential record, the interaction among heterogeneous events could be complex, it is difficult to capture the dynamic interaction events representation. (2). The total dimensional numbers of the heterogeneous events embedding could be disastrously vast, making it impossible for sequential model to capture the temporal representation.

In order to effectively capture the dynamic heterogeneous temporal events representation, we propose attentional multihead aggregation.Given a time step events embedding $E_t \in R^{40,d}$, M randomly initialized head-masks are generated. Each mask is used to capture different aspects of events-interaction at $time_t$ with attention-based modules. At last, all heads are concatenated to produce the eventual aggregation representation.To be specific, the details of the aggregation modules are shown as following:

$$H_i = \sum_{j=1}^{40} attscore_i(K_j, Mask_i) \cdot (W_{vi} E_j) \quad (5)$$

$$attscore_i(A, B) = Softmax(attention_i(A, B)) \quad (6)$$

$$attention_i(A, B) = (W_{ki} A) \cdot B \quad (7)$$

$$agr = Concat(H_1, H_2, ..., H_m) \quad (8)$$

Where $H_i \in R^d$ is the i^{th} head of aggregation representation, $agr \in R^{M \cdot d}$ concatenates all heads , $attscore_i$ is the attention mechanism of the i^{th} Head, getting the aggregation proportion of each event with calculating dot product between events and Mask. Each head could capture its own concerned events information with its corresponding formed transform matrices W_k,W_v and Mask. Dynamic heterogeneous events interaction could be implicitly captured during multihead aggregation.Example of two-heads aggregation is shown as figure 2

3.3. Sequential Model and Prediction

: For each time step, we get an aggregation representation.Given a sample $record \in R^{L,40}$, a temporal events aggregation representation $A \in R^{L,M \cdot d}$ is captured. As the Long Short-Term Memory (LSTM) be successfully used in sequential data, we pass on A to one-layer bidirectional LSTM module.We sum up the last-time outputs of both forward and backward units and get the logits through a single dense layer with sigmoid activation.
Our objective is a binary classification. The loss function used is:

$$L(y, \hat{y}) = -(y \cdot \hat{y}) + ((1 - y) \cdot (1 - \hat{y})) \quad (9)$$

4. Experiment Results

Table 1. Test-metrics for different model

Model	AUROC	AUPRC	Utility score
MLP	0.7763	0.0771	0.2413
Dense-Embedding	0.8013	0.0923	0.3679
8-heads-HEA-Transformer	0.8264	0.1236	0.3968
8-heads-HEA-LSTM	0.8317	0.1289	0.4023
16-heads-HEA-LSTM	0.8354	0.1307	0.4126

The provided data is divided into train and test sets at the ratio of 7 to 3 respectively over sepsis and normal patients.To do a further experiment, we divide the data into train and host sets at the ratio of 9 to 1.And then 5 cross validation sets are averagely partitioned with the train sets.Evaluation metrics are measured over the host set. To evaluate our proposed aggregation modules, we directly use a MLP prediction model, utilize a single dense layer as embedding, replace LSTM with transformer, set various numbers of heads to train different models. The results show that our proposed model obviously improve the metrics, and have high efficiency.

We measure Area Under the Receiver Operating Characteristic curve(AUC) and Area Under The Precision Recall Curve(APC) as our evaluation metrics.

Dense Representation Layer: Given the sampled record X, the dense representation layer use a single dense layer with activation to generate the temporal representation $\hat{x} \in R^{L,40}$

$$\hat{x} = xW + b \quad (10)$$

$W \in R^{40,d}$ is the transformation matrices, $\hat{x} \in R^{L,d}$.

Table 2. Hold-metrics for HETA

Model	AUROC	AUPRC	Utility score(average/major vote/any vote)
8-heads-HEA-Transformer	0.8186	0.1049	0.3641/0.3635/0.3658
8-heads-HEA-LSTM	0.8206	0.1031	0.3656/0.3613/0.3643
16-heads-HEA-LSTM	0.8224	0.1052	0.3817/0.3750/0.3830

Table 3. Training time of different models

Model	Time(*\ min)
8-heads-HEA-Transformer	4
8-heads-HEA-LSTM	9
16-heads-HEA-LSTM	10

Transformer:Given a generated temporal events aggregation representation $A \in RL, d$, transformer use position-encoding and multihead self-attention modules to capture the temporal information. Via a multilayer time-wise self-attention module, temporal information will be captured within the representation $\hat{A} \in R^{L,d}$.At last, a single dense layer with sigmoid activation is used to get logits within the reshape of $\hat{A}$.

Multihead aggregation: We keep dimensional numbers(d) as 16 and set heads to be 8 and 16 as 2 different aggregation modules. Our result shows that 16 heads get the best metrics.We firstly do experiment for different models over train and test sets. The results over test set are showed as following Table 1.

The result shows that heterogeneous temporal aggregation modules could improve the metrics obviously,then we do the further experiment over cross validation sets. The results over hold set are showed as following Table 2.

What's more, our proposed model is in high efficiency, as it just need to calculate the attention score between events and multiheads.The aggregation timely cost for 1 time step is O(Ed^2M), where E is the numbers of events, d is the dimensional numbers and M is the numbers of heads. With one GPU, the experimental training time of different models over 1024000 samples is showed as following Table 3.

We have run the official scoring code with 16-heads-HETA-LSTM and got the following results of different metrics: AUROC (0.866), AUPRC (0.293), Accuracy (0.89), and Utility (0.402).

5. Conclusion

We proposed an attention-based sequential representation model to do early sepsis prediction from clinical data. Our proposed model includes two main parts: clinical events co-occurrence with heterogeneous event aggregation and temporal interaction capture with LSTM. Experiments in the PhysioNet/Computing in Cardiology Challenge 2019 show that the heterogeneous event aggregation module can shorten the length of clinical event sequence for better temporal dependency modeling, and the separated storage strategy of aggregation representation with different heads retains temporal interactions of events.

Acknowledgments

This paper is partially supported by Beijing Municipal Commission of Science and Technology under Grant No. Z181100008918005, National Key Research and Development Program of China with Grant No. SQ2018AAA010010, and the National Natural Science Foundation of China (NSFC Grant No. 61772039 and No. 91646202).

References

[1] Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics 2017;bbx044.

[2] Qian F, Gong C, Liu Lc, Sha L, Zhang M. Topic medical concept embedding: Multi-sense representation learning for medical concept. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2017; 404–409.

[3] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation 1997;9(8):1735–1780.

[4] Che Z, Purushotham S, Li G, Jiang B, Liu Y. Hierarchical deep generative models for multi-rate multivariate time series. In International Conference on Machine Learning. 2018; 783–792.

[5] Liu L, Li H, Hu Z, Shi H, Wang Z, Tang J, Zhang M. Learning hierarchical representations of electronic health records for clinical outcome prediction. AMIA Annual Symposium 2019;.

[6] Liu L, Shen J, Zhang M, Wang Z, Tang J. Learning the joint representation of heterogeneous temporal events for clinical endpoint prediction. In Thirty-Second AAAI Conference on Artificial Intelligence. 2018; .

Address for correspondence:

Ming Zhang
1628, No.1 Science Building, Peking University, Beijing, China
mzhang_cs@pku.edu.cn