ISSN: 2664-9640

Non-Uniformly Spaced Time Series Prediction Based on LSTM and Attention Mechanism

Xiaolong Xu

College of Computer Science, Sichuan University, Sichuan, Chengdu, China.

Abstract

Time series prediction tasks are widely used, such as stock price prediction in finance, cargo arrival time prediction in logistics, and vehicle speed prediction in transportation, which are very important for decision makers. In these tasks, LSTM model has played an important role in time series prediction. However, traditional LSTM models usually assume uniform time intervals in time series prediction, and datasets with non-uniform time intervals pose a new challenge to LSTM. In this paper, we construct a dataset with non-uniform time intervals based on the dataset "Evolução Diária dos Acionamentos de Meios de Emergência Médica", which is a dataset of the daily use of medical emergency vehicles in Portugal, and construct an Attention LSTM model and a Multi-Headed Attention LSTM model with optimisation of the attention mechanism. The Attention-LSTM model and the Multi-Headed-Attention-LSTM model are constructed by optimising the LSTM with attention mechanism. The results show that our model can be well adapted to the non-equal time interval data and has a wide range of application prospects.

Keywords

Non-Uniformly spaced time series, LSTM, Attention mechanism, Multi-head Attention mechanism.

1. Introduction

In the real world, a lot of time-series data are non-equally spaced, such as stock prices in the financial sector, freight data in the logistics sector, and vehicle travel speeds in the transportation sector. Predictions of such data are very important for decision makers because they can help people make more informed decisions and help decision makers optimize resource allocation. Past research has focused on the prediction of equally spaced time-series data, such as methods based on traditional LSTM[1] models. However, these methods have limitations when dealing with non-equally spaced time-series data because they cannot deal with the irregularity of time intervals.

In order to solve the shortcomings of traditional LSTM on non-equally spaced time series data, we construct an improved LSTM model based on the processed "*Evolução Diária dos Acionamentos de Meios de Emergência Médica*"[4] dataset, which is capable of accomplishing the task of equally spaced time series prediction very well. This model is able to fulfill the task of equal-interval time series prediction very well. In order to make the non-isochronous temporal prediction task available. We transform the temporal information of the dataset into input features to realize the prediction of the data after an arbitrary time interval.

The experimental results show that for the prediction of non-equally spaced temporal data, both our constructed Attention-LSTM[5] as well as Multi-Headed-Attention-LSTM can fulfill the task well. In the same case of using the transformation of temporal information into input features, Multi-Headed-Attention-LSTM performs better than Attention-LSTM, which shows that for our dataset "Evolução Diária dos Acionamentos de Meios de Emergê ncia Médica", we

ISSN: 2664-9640

not only need to consider the historical important data, but also the correlations between features should be fully utilized.

The contributions of this paper are as follows:

In this paper, we have constructed a variant of LSTM model based on non-equally spaced time series data prediction using the attention mechanism[2], which can be well adapted to the task of non-equally spaced time series data prediction, and has a wide range of application prospects.

In this paper, we processed the "*Evolução Diária dos Acionamentos de Meios de Emergência Médica*" dataset to construct non-equally spaced time-series data, and verified the validity of the constructed models using the data.

Specifically , the paper is divided into several sections. The second section covers the related work, while the third section provides a comprehensive account of the data processing and model construction. Appropriate metrics for training the model are also selected in this section. The fourth section validates the model's effectiveness by training it using the processed dataset and presenting the prediction results on the test set. Finally, the last section concludes the entire work.

2. Related Work

Forecasting with non-uniformly spaced time series, where time intervals between data points are irregular, is a difficult task that has garnered significant attention from researchers in recent years. To address this challenge, several methods have been proposed, including models like T-LSTM(Time-Aware Long-Short Term Memory)[6] and ATTAIN[7]. These approaches aim to improve the accuracy and efficiency of forecasting by accounting for the non-uniformity of the time intervals in the time series data.

2.1. Time-Aware Long-Short Term Memory

The T-LSTM model is an improved version of the LSTM model. It takes into account the effect of time on prediction by adding time weights and adapts the model's memory and forgetting mechanism to handle non-equally spaced time series data by weighting importance.

The main idea of the T-LSTM model is to assign different weights to each time step, allowing the model to focus more on the most recent time step, thus capturing more detail in the time series data. The key to the model is the calculation of the time weights, which takes into account the temporal distance between each time step and the prediction target, with closer time steps having higher weights.

2.2. ATTAIN

The ATTAIN model was proposed in 2019 by Zhang, Y.[7] et al. They proposed a model called ATTAIN for modelling disease progression based on patient history. The core idea of the ATTAIN model is to combine TA-LSTM with an attention mechanism to better handle non-equally spaced time series data and to automatically select and focus on the most relevant historical records.

3. Methods

3.1. Data Processing

3.1.1. Data Preparation

The dataset *"Evolução Diária dos Acionamentos de Meios de Emergência Médica"* is a count of daily hospital emergency transport equipment in Portugal, as shown in Figure 1. The raw data are continuous time series, which need to be processed in order to construct non-equal time series for the prediction of the target.

ISSN: 2664-9640

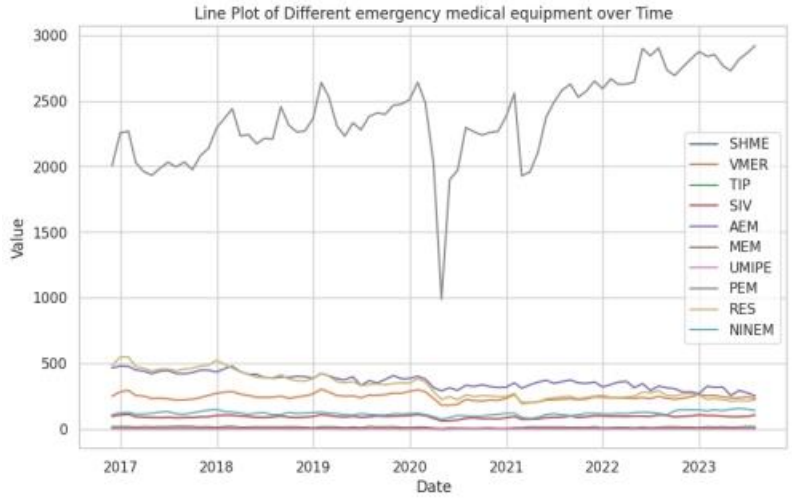

Figure1. Different emergency medical equipment

3.1.2. Lone Forest Outlier Detection

For detecting outliers we have used lone forest outlier detection respectively. Lone forest outlier detection is a random forest based outlier detection method. Its basic principle is to randomly divide the dataset into many subsets and train each subset using the random forest model. Then, for a new data point, it is predicted by the random forest model and the average distance between it and other data points is calculated. If this distance is greater than a certain threshold, it is considered an outlier.

The advantage of lone forest outlier detection is that it can handle high-dimensional datasets, is appropriate for processing our data, and does not require prior normalisation or standardisation of the data. Some of the results of continued outlier detection using lone forest are shown in Figure 2.

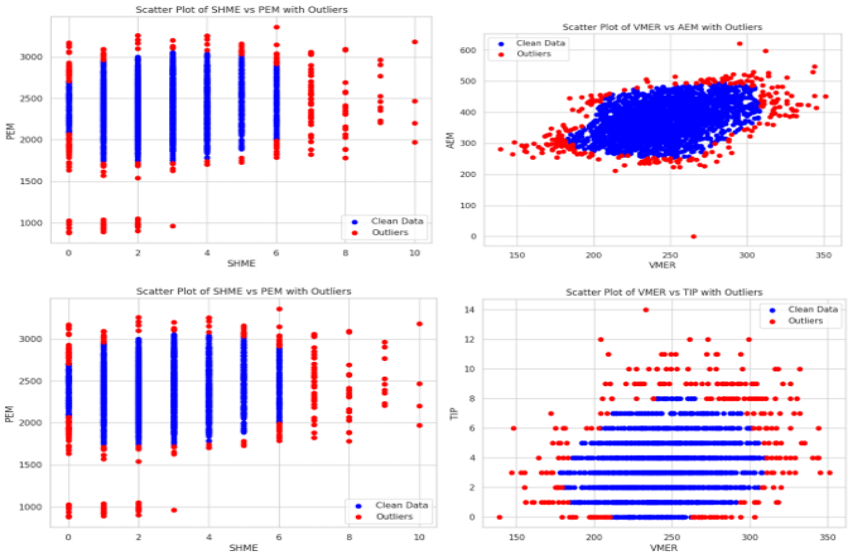

Figure 2. Results of the Lone Forest outlier test

We believe that the deletion of these outliers is beneficial to the construction of non-equally spaced time series data, and the deletion of the data not only removes the outliers, but also achieves the construction of non-equally spaced time series data, which provides reliable data for the subsequent construction and validation of the model. In addition, we still need to consider the continuous part of the data, we take the way of sampling from the starting time of different time spacing for further processing of the dataset. Finally, we constructed our non-equally spaced time series dataset.

ISSN: 2664-9640

3.1.3. Data Representation:

After our processing, the dataset can be represented as $D = \{d_1, d_2, d_3 \dots \dots d_M\}$,where M denotes the total number of data samples, and for each d_i, it consists of n features, $d_i = \{d_i^1, d_i^2, d_i^3 \dots \dots d_i^n, b_t\}$, where P denotes the number of features in the data sample, d_i^kdenotes the value of the k-th feature in a record,b_t denotes the time interval from the previous data record in a non-uniform time interval record. Our task is to predict the output at the time interval from the current input given the data set of the previous record , and the time interval of the record . Our task is to predict the output d_{t+pt} at the moment p_t for the current input time $t$interval, given the dataset $D_{t-1} = \{d_1, d_2, d_3, \dots \dots d_{t-1}\}$ of the first $t-1$ records, and the record interval p_t.

3.2. Models

3.2.1. Attention-LSTM

The conventional LSTM model is designed to work with time series data that is continuous and uniform in terms of time intervals. However, the introduction of the attention mechanism can help to assign different weights to various elements in the sequence data dynamically. This enables the model to better concentrate on the essential aspects of the input data, thereby enhancing the model's predictive ability. The pseudo code for its model training implementation is as follows:

Algorithm1: Attention-LSTM

1: ***Input****: Training Set(X, y), Number of iterations* $\boldsymbol{T}$*, Learning rate* $\boldsymbol{\eta}$*, Loss function* $\boldsymbol{L}(\boldsymbol{y}, \hat{\boldsymbol{y}})$

2: ***Output:*** $\hat{\boldsymbol{y}}$

3: For i=0 to T:

4: ***forward propagation****:*

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right)$$
$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$
$$\tilde{c}_t = tanh(W_c x_t + U_c h_{t-1} + b_c)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$
$$h_t = o_t \odot tanh(c_t)$$
$$a_t = softmax(v^T tanh(W_a h_t + b_a))$$

5: $$context_t = \sum_{i=1}^{T} a_{t,i} h_i$$

6: $$y_t = W_y context_t + b_y$$

7: ***loss calculation****:* $loss = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{M}\left(y_{ij} - \widehat{y_{ij}}\right)^2$

8: ***backpropagation****:*

$$\frac{\partial loss_t}{\partial context_t} = \frac{\partial loss_t}{\partial y_t} W_y^T \frac{\partial y_t}{\partial context_t} = \left(y_t - y_{true,t}\right) \cdot W_y^T$$
$$\frac{\partial loss_t}{\partial \alpha_{t,i}} = \frac{\partial loss_t}{\partial y_t} W_y^T \frac{\partial context_t}{\partial \alpha_{t,i}} = \left(y_t - y_{true,t}\right) \cdot W_y^T \cdot h_i$$
$$\frac{\partial loss_t}{\partial h_{t-1}} = \frac{\partial loss_t}{\partial h_t} \cdot o_t \cdot (1 - tanh^2(c_t)) \cdot U_f + \frac{\partial loss_t}{\partial c_t} \cdot f_t \cdot (1 - tanh^2(c_{t-1}))$$
$$\frac{\partial loss_t}{\partial c_{t-1}} = \frac{\partial loss_t}{\partial h_t} \cdot o_t \cdot (1 - tanh^2(c_t)) \cdot f_t + \frac{\partial loss_t}{\partial c_t} \cdot f_t \cdot (1 - tanh^2(c_{t-1}))$$

11: ***gradient update****:* $\theta \Leftarrow \theta - \eta \frac{\hat{m}}{\sqrt{\hat{v}} + \epsilon}$

12: End For

3.2.2. Multi-Headed-Attention-LSTM

The advantage of the multi-head attention[3] mechanism over the single-head attention mechanism is that it is able to learn more semantic information, which improves the expressiveness of the model. While the single-head attention mechanism can only focus on one

ISSN: 2664-9640

part of the input sequence, the multi-head attention mechanism can focus on several parts of the input sequence at the same time, thus better capturing the key information in the sequence. In addition, the multi-head attention mechanism can also improve the robustness of the model as it can learn different attention weights, thus reducing the model's dependence on some specific parts of the input sequence. The pseudocode for the model training implementation is shown below:

Algorithm2: Multi-Headed-Attention-LSTM

1: ***Input***: *Training Set*(X, y), *Number of iterations* $\boldsymbol{T}$, *Learning rate* $\boldsymbol{\eta}$, *Loss function* $\boldsymbol{L}(\boldsymbol{y}, \widehat{\boldsymbol{y}})$, *attention head numbers* $\boldsymbol{H_{att}}$

2: ***Output:*** $\widehat{\boldsymbol{y}}$

3: *For i=0 to T:*

4: ***forward propagation***:

5: $h_t^l, c_t^l = \boldsymbol{LSTM}\left(x_t, h_{t-1}^l, c_{t-1}^l\right)$;

6: $\boldsymbol{head_h} = \boldsymbol{softmax}\left(\frac{\mathbf{1}}{\sqrt{\boldsymbol{H}}}\boldsymbol{q_h k_h^T}\right)\boldsymbol{v_h}$

7: $att = [head_1, head_2, head_3 \ldots\ldots head_{H_{att}}]$

8: $y = \mathbf{W}_{att} + \boldsymbol{b}$

9: ***loss calculation***: $loss = \frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij} - \widehat{y_{ij}}\right)^2$

10: ***backpropagation***: $\frac{\partial \boldsymbol{loss}}{\partial \boldsymbol{P_l}} = \frac{\partial \boldsymbol{loss}}{\partial \boldsymbol{y}} \cdot \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{att}} \cdot \frac{\partial \boldsymbol{att}}{\partial \boldsymbol{P_l}}$;

$$\frac{\partial \boldsymbol{att}}{\partial \boldsymbol{head}} = [\frac{\partial \boldsymbol{head1}}{\partial \boldsymbol{head}}, \frac{\partial \boldsymbol{head2}}{\partial \boldsymbol{head}}, \frac{\partial \boldsymbol{head3}}{\partial \boldsymbol{head}}, \ldots\ldots \frac{\partial \boldsymbol{head_{H_{att}}}}{\partial \boldsymbol{head}}]$$

$$\frac{\partial \boldsymbol{head_h}}{\partial \boldsymbol{W_{q,h}}} = \boldsymbol{x_T}\left(\frac{\partial \boldsymbol{head_h}}{\partial \boldsymbol{q_h}} \cdot \frac{\partial \boldsymbol{q_h}}{\partial \boldsymbol{W_{q,h}}}\right);$$

$$\frac{\partial \boldsymbol{head_h}}{\partial \boldsymbol{W_{k,h}}} = \boldsymbol{x_T}\left(\frac{\partial \boldsymbol{head_h}}{\partial \boldsymbol{k_h}} \cdot \frac{\partial \boldsymbol{k_h}}{\partial \boldsymbol{W_{k,h}}}\right);$$

$$\frac{\partial \boldsymbol{head_h}}{\partial \boldsymbol{W_{v,h}}} = \boldsymbol{x_T}\left(\frac{\partial \boldsymbol{head_h}}{\partial \boldsymbol{v_h}} \cdot \frac{\partial \boldsymbol{v_h}}{\partial \boldsymbol{W_{v,h}}}\right);$$

11: ***gradient update***: $\theta \Leftarrow \boldsymbol{\theta} - \boldsymbol{\eta}\frac{\widehat{\boldsymbol{m}}}{\sqrt{\widehat{\boldsymbol{v}}} + \boldsymbol{\epsilon}}$

12: *End For*

3.3. Choice of Comparison Metrics and Hypothesis Tests

In a multi-output regression prediction task, we can use MSE, $RMSE$ and $R2 - Score$ metrics to evaluate the performance of the model.

MSE is the mean square error, an important assessment metric that measures the accuracy of a model's predictions. It calculates the difference between the predicted value and the true value, squares the difference, and then averages the values.The smaller the MSE, the smaller the model's error with respect to the predicted outcome, which means the model's predictive ability is superior.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij} - \widehat{y_{ij}}\right)^2$$

Where N denotes the number of samples, M denotes the number of output features, y_{ij} denotes the true value of the j-th output feature of the i-th sample, and $\widehat{y_{ij}}$ denotes the predicted value of the j-th output feature of the i-th sample.

MAE is the mean square absolute error, which is used to measure the average absolute error between the predicted and true values of a model. Compared to MSE, the calculation of MAE does not involve squaring the error, thus maintaining the consistency between the measure and the original value and avoiding the problem of error amplification that may be caused by the

ISSN: 2664-9640

squaring operation. In addition, MAE has better robustness to outliers because it has no squaring term to amplify the effect of outliers.The calculation results of MAE are easier to understand and interpret, and are usually expressed in terms of units of predicted values. Its formula is as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left|y_{ij} - \widehat{y_{iJ}}\right|$$

Root Mean Squared Error (RMSE) is the square root of the mean square error.RMSE is similar to MSE in that it requires a squaring operation for MSE.RMSE is calculated as follows

$$loss = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij} - \widehat{y_{iJ}}\right)^2}$$

R2-Score (Coefficient of Determination) is a measure of how well the model fits the data.R2-Score can take values from 0 to 1, with the closer the value to 1, the better the model fits the data. The formula for calculating R2-Score is as follows

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij} - \widehat{y_{ij}}\right)^2}{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij} - \overline{y_j}\right)^2}$$

where $\overline{y_j}$ denotes the mean of the j-th output feature

3.4. Model Training

For our processed dataset, we used the constructed Attention-LSTM model and the Multi-Headed-Attention-LSTM model for the prediction task. The loss profile of the model training is shown in Figure 3.

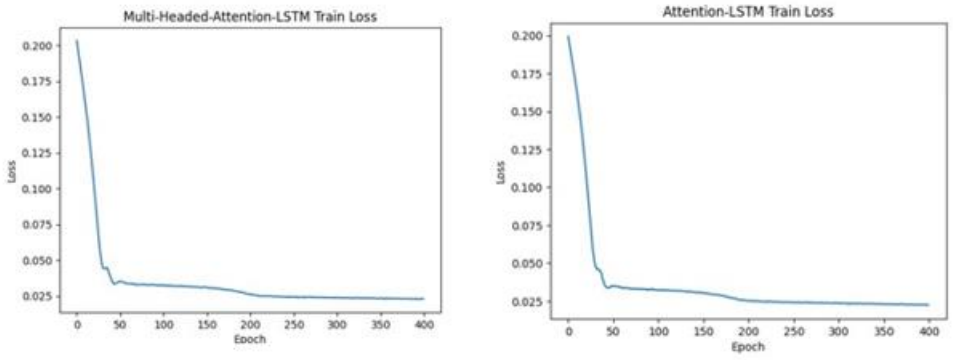

Figure 3. Training loss plot

From Figure 3, we can see that after our training rounds reach 250, the training loss gradually tends to stabilise and we consider that the model has reached the optimal prediction.

4. Results

In order to verify the effectiveness of our constructed Attention LSTM as well as the Multi-Headed Attention LSTM on non-equally spaced temporal data, 80% of the processed dataset "*Evolução Diária dos Acionamentos de Meios de Emergência Médica*" is used as the training dataset and 20% of the test dataset is used to train and predict the model, and the model's data for the different metrics on the test set are shown in Table 1.

ISSN: 2664-9640

Table 1: Performance of two Models

Model	MAE	MSE	R2
Attention-LSTM	33.9413	78.8471	0.9814
Multi-Headed-Attention-LSTM	32.1047	75.4828	0.9914

We find that both Attention-LSTM and Multi-Headed-Attention-LSTM perform better, with R2 scores greater than 0.98, when temporal information is simultaneously converted to input features. However, Multi-Headed-Attention-LSTM achieves better results compared to Attention-LSTM because the simultaneous attention to different features through multiple attention computations makes the model better able to capture the relationship between features and achieve better results.

5. Conclusion

In this study, we processed the "Evolução Diária dos Acionamentos de Meios de Emergência Médica" dataset to obtain non-equally spaced time series data. Subsequently, we developed two improved LSTM models based on the attention mechanism: the Attention LSTM model and the Multi-Headed Attention LSTM model. These models were utilized to predict the non-equally spaced time series data using the processed dataset. After training the models on the processed dataset, we achieved promising results. Our findings suggest that the improved LSTM models based on the attention mechanism are efficacious in predicting non-equally spaced time series data. The model constructed in this paper achieves good performance in predicting non-equally spaced time series and has a wide range of applications.

References

[1] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), 2015.

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[4] SNS. (2021). Evolução Diária dos Acionamentos de Meios de Emergência Médica. https://transparencia.sns.gov.pt/explore/dataset/acionamentos-de-meios-de-emergencia-medica/

[5] Kaihua Tang, Xiaohui Yan, Yuxuan Lai, and Qiang Yang. Attention-based recurrent neural network models for joint intent detection and slot filling. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1558–1568, 2016.

[6] Michigan State University, IBM Research, and Cornell University. (2018). Time aware long short-term memory. Knowledge Discovery and Data Mining (KDD) conference.

[7] Zhang, Y. and Gan, Z. (2019). ATTAIN: Attention-based Time-Aware LSTM Networks for Disease Progression Modeling. Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), 116-123.