ISSN: 2664-9640

Covid-19 prediction model based on GRU-LSTM

Yanmei Zhao[1], Kun Wei [2,*]

[1] Department of Neonatology, Henan (Zhengzhou) Hospital (Children's Affiliated to Zhengzhou University, China

[2]College of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, Henan 450000, China

* Corresponding author: Kun Wei

Abstract

The worldwide pandemic of COVID-19 has had a serious impact on the whole society. Predicting the number of confirmed cases through mathematical modeling will help to provide a basis for public health decision-making. In the complex and changeable external environment, the infectious disease prediction model based on deep learning has become a research hotspot. However, the existing models have high requirements for the amount of data, and can 't adapt well to the scenario with low amount of data during supervised learning, resulting in the reduction of prediction accuracy. Aiming at the problems of low prediction accuracy and long prediction time of current load prediction models, a combined prediction model GRU-LSTM based on gating cycle unit (GRU) and short-term memory (LSTM) network is established. The network structure of the model includes three layers. The first layer adopts GRU, which reduces the training time of the model by using the characteristics of few GRU parameters and easy convergence. The second and third layers adopt LSTM, which improves the prediction accuracy of the model by combining the advantages of many LSTM parameters. On this basis, the data set is processed by missing values and standardization, and a new set of sequence values is obtained after feature selection of the original sequence using random forest algorithm. The sequence values are used as the input of GRU-LSTM combined prediction model. The experimental results show that the hybrid model can effectively improve the prediction performance, compared with convolution neural network, cyclic neural network, long-term and short-term memory network and gated cyclic unit model, GRU-LSTM model performs well in the evaluation indicators of mean absolute percentage error and root mean square error, and is more suitable for predicting COVID-19 propagation trend.

Keywords

COVID-19; LSTM; GRU; Machine learning.

1. Introduction

At the end of 2019, COVID-19gradually evolved from a local outbreak to a worldwide pandemic. As of December 23, 2020, COVID-19 has infected more than 78million people worldwide and killed more than 1.7 million people in total. Accurate prediction of the number of confirmed cases can help decision makers formulate epidemic prevention and control measures and long short-term treatment plans and measures, which is of great significance for the effective control of the epidemic [1,2].

Researchers have proposed a large number of mathematical models to model and predict the diffusion and propagation trend of COVID-19. The classical epidemic model needs manual subsection design and parameter estimation in complex and changeable situations (such as

ISSN: 2664-9640

changes in policies and external conditions), which is inflexible and ineffective. The composite population model has very high requirements for migration data between populations. The defect of any original data will cause the prediction error of the model to become larger, and it is difficult to model in the scene of large samples. Therefore, in the complex and changeable environment, the infectious disease transmission model based on deep learning has gradually become a research hotspot. However, at present, there is a serious shortage of data in COVID-19 research. The existing supervised learning methods cannot well adapt to low data scenarios, and the prediction accuracy of the model is low [3].

Coronavirus has great similarity in biological characteristics, with very similar disease manifestations, transmission routes and development trends. Taking this characteristic as a modeling factor can make the prediction model learn the relevant characteristic information of the virus in advance and effectively assist in the prediction of the trend of confirmed cases. This paper constructs COVID-19 prediction model GRU-LSTM based on pre training fine tuning strategy. Using the pre training strategy on the existing data set, the model can be exposed to more epidemic data in advance, so as to obtain more sufficient prior knowledge. At the same time, the impact of local artificial restriction policies on epidemic trends is taken into account in the model to achieve accurate prediction in the data set of target areas.

2. Related work

In terms of COVID-19 transmission prediction, classical infectious disease transmission models predict the trend of COVID-19 transmission through mathematical modeling, such as Sir infectious disease model, improved SEIR model, etc. The prediction model based on deep learning learns low dimensional features through multi-layer nonlinear structure to form a more abstract high-dimensional representation, which has strong expression ability. Chimmula et al. Used the LSTM network to predict the end date of the epidemic in Canada [4]. The short-term accuracy of the model is 93.4%, and the long-term accuracy is 92.67%. Arora et al. Used LSTM and its variants to predict the number of positive cases in India. The daily prediction error of this method is less than 3%, and the weekly prediction error is less than 8%. Although using LSTM can better predict the overall trend of the number of confirmed cases, LSTM is not sensitive to the change of a certain parameter. For example, it is difficult to effectively predict the surge in the number of confirmed cases in a certain period of time due to the implementation of national policies. In addition, Huang et al. Proposed using convolutional neural network (CNN) to analyze and predict the number of confirmed cases. However, the above-mentioned in-depth learning methods do not take into account the impact of complex and volatile factors on the epidemic. Considering the impact of some external factors on the epidemic, Yang et al. Combined with social and economic characteristics, based on the gated recurrent unit (GRU), studied the epidemic data and epidemic time series in the United States, and then predicted the future epidemic transmission trend However, supervised learning requires a high amount of data, and insufficient data will lead to poor prediction effect of the model.

As a variant of LSTM, GRU shortens the prediction time by reducing one gate when predicting time series. Guo et al proposed a multi-step prediction method combining GRU and autocorrelation analysis. The simulation results show that the prediction time of the proposed GRU prediction model is optimized during load prediction. For some specific timeseries problems, using temporal convolutional network (TCN) to model can also achieve good results. Literature [5] used TCN to conduct experiments on several real-world data sets. The results show that TCN has good effects on point prediction and probability prediction, but poor effects on long-term prediction.

ISSN: 2664-9640

The existing prediction models are usually single prediction models or some combined prediction models based on integrated learning [6]. Although the single prediction model is better than the combined prediction model in prediction time, the prediction accuracy is significantly lower than the combined prediction model. Although the prediction accuracy of some existing combined prediction models is higher than that of a single prediction model, the prediction accuracy is relatively low and the problem of prediction time is not considered. In order to improve the prediction accuracy and efficiency, combined with the respective prediction advantages of GRU and LSTM, this paper proposes a combined prediction model based on GRU and LSTM to predict COVID-19.

3. System Components

This paper constructs COVID-19 prediction model GRU-LSTM based on pre training fine tuning strategy. Through the pre training strategy, the problem that the accuracy of the prediction model is reduced due to the insufficient amount of data is solved to a certain extent, and richer initialization parameters are provided for the prediction model, so that the model can learn the essential law of COVID-19 virus in advance. After fine tuning on the new data set, it has higher prediction accuracy for the development trend of confirmed cases. The load condition at each time is closely related to the load condition at the previous time. The closer the historical value is to the value of t at the current time, the closer the relationship between them is. For model selection, long-distance dependent information can provide trend information, which cannot be completely ignored. In order to make better use of past load data, historical data needs to be selectively retained and discarded. The forgetting gate in LSTM can control the amount of historical information entering the current time and the amount of information that needs to be discarded. Therefore, the model in this paper selects LSTM and GRU with similar network structure to LSTM

3.1. LSTM model

As an improved RNN, LSTM [7] inherits the advantages of RNN model, and effectively solves the problems of gradient explosion and gradient disappearance in RNN by using the unique gate structure. Therefore, LSTM can effectively deal with long-time series problems, and has been successfully applied to speech recognition, image description, natural language processing and other fields. Compared with RNN and GRU, the fitting and prediction accuracy of LSTM model is generally higher, but the training process is time-consuming due to too many LSTM parameters.

LSTM is composed of multiple cycle units. By updating neuron state information, input gate, output gate and forgetting gate are used to control the weight of historical information, so as to store past information.

3.2. GRU model

GRU [8], as a variant of LSTM, can also effectively solve the problems of gradient explosion and gradient disappearance in RNN. Compared with LSTM, GRU has a simpler structure, which combines forgetting gate and input gate into an update gate. Since GRU reduces one gate and matrix multiplication becomes less, a lot of time can be saved when the amount of training data is large.

3.3. GRU-LSTM model

Considering the two factors of prediction accuracy and prediction time, this paper establishes a combined prediction model based on GRU and LSTM to predict COVID-19. The network structure of the combined prediction model is shown in Figure 1.

ISSN: 2664-9640

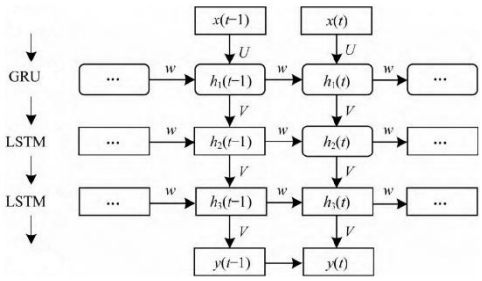

Figure.1 GRU-LSTM model

The network structure of the combined prediction model in this paper includes three layers: the first layer adopts GRU, because the network structure of GRU is simple, the parameters are less, and it is easier to converge, so the GRU training speed is fast when training data, which can reduce the training time, but the prediction accuracy of GRU is lower than LSTM; The second and third layer network structures adopt LSTM, which has more LSTM parameters and higher prediction accuracy, and the prediction accuracy of double-layer LSTM is better than that of single-layer LSTM.

3.4. Model evaluation criteria

In this paper, mean absolute error (MAE), mean absolute percentage error (MAPE), mean square error (MSE), root mean square error (RMSE) and coefficient of determination (R2) are used as the performance evaluation criteria of the model. The prediction time mainly takes the training time of the model as the evaluation standard. In the generalization experiment of the model, the explained variance score is used to evaluate the fitting degree of the model.

4. Experiments

4.1. Experiment environment

Experimental environment of this research was shown in Table 1.

Table 1.Experimental environment.

Environment	Value
OS	Windows10
CPU	I5-12400 H
Memory	32G
Language	Python
Tool	Anaconda-Keras

4.2. Data preprocessing

This experiment uses Google cloud platform data set, which contains daily time series data related to COVID-19, distributed in 20000 different locations around the world. The data period used in the experiment is from January 1, 2020 to November 26, 2020. In the experiment, data from India and the United States were selected, pre trained on the Indian data set, and fine tuned on the American data set. The influencing factors of characteristic data as input include static characteristic data and dynamic characteristic data. The influencing factors of static characteristic data include local per capita GDP, demographic data, average life expectancy of local people, etc., while the influencing factors of dynamic characteristic data include COVID-19 case data (daily infection cases, cumulative infection cases, death data), government intervention policy data, etc. At the same time, there are many other factors related to the spread of the epidemic.

ISSN: 2664-9640

4.3. Experimental Performance Evaluation

After obtaining the data, first, the missing value and standardization of the original data are processed. When the missing rate of the original data is greater than 30%, the data is discarded. If the missing rate is less than or equal to 30%, the missing value is filled with the mean filling method, and the filled data is standardized through the normalization method. The random forest algorithm is used to select the features of the standardized data, according to the importance of features, add the feature data obtained after feature selection to the weight parameters for combination, input the combined load value into the GRU-LSTM combined prediction model for training, and set the step size to 12 (predict the 13th data according to the first 12 data). Finally, five evaluation criteria are used to evaluate the performance of the model, and the prediction results of the model are output at the same time. The parameter settings of the prediction model are shown in Table 2.

Table 2. Experimental training parameters

Training parameters	Parameters setting
GRU units	24
First LSTM layer	24
Second LSTM layer	24
Active function	tanh
Iterations	400

By comparing with the current mainstream models, we can objectively and fully verify the effectiveness of this model. CNN model, recurrent neural network (RNN) model, LSTM network model, GRU model and GRU-LSTM model without pre training strategy (no-pretrain-GRU-LSTM) were selected as the comparison models in the experiment as table 3

Table 3. Experimental Results

Category	MAPE/%	RMSE/%
CNN	18.45	6.8
RNN	21.21	2.9
GRU	20.98	2.4
LSTM	2.7	4.6
GRU-LSTM	1.5	1.4

5. Conclusion

Aiming at the problems of low prediction accuracy and long prediction time of COVID-19 prediction models, this paper proposes a combined prediction model GRU-LSTM based on LSTM and GRU. This paper presents a COVID-19 prediction model GRU-LSTM combined with pre training fine tuning strategy, and compares the performance of root mean square error and average absolute percentage with CNN, RNN and other models on the data set in the United States. The experimental results show that this model based on supervised learning can solve the problem of insufficient accuracy of the model caused by the small amount of data to a certain extent, and help to improve the performance of trend prediction of confirmed cases. For the transmission of mutated novel coronavirus, the pre training fine tuning strategy proposed in this paper can also be used to predict the transmission trend of the epidemic. In the next step, we will analyze the impact of exogenous factors on COVID-19 transmission, and add other factors related to epidemic transmission as features to the model, such as the prevalence of masks, people's awareness of protection.

ISSN: 2664-9640

References

[1] Zhou Tao, Liu Quanhui, Yang Zimo, Liao Jingyi, Yang Kexin, Bai Wei, Lu Xin, Zhang Wei Preliminary prediction of basic regeneration number of novel coronavirus pneumonia [j] Chinese Journal of evidence based medicine, 2020,20 (03): 359-364.

[2] Ji Hanran, Wu Jiewen, Yang Xinping, Pang Mingfan, Zhao Qing, Liang zuoru, Fang Yuansheng, Zhang rongna, Li Jie, Qi Xiaopeng Risk assessment of global covid-19 epidemic in March 2022 [j] Disease surveillance, 2022,37 (04): 430-434Covid-19 propagation modeling and prediction in urban rail transit system [j] Lei bin, Liu Xingliang, Cao Zhen, Hao Yarui, Zhang Yuan, Chen Xinmiao Journal of transportation engineering 2020(03).

[3] Time series forecasting of COVID-19 transmission in Canada using LSTM networks[J] . Vinay Kumar Reddy Chimmula,Lei Zhang. Chaos, Solitons and Fractals: the interdisciplinary journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena . 2020 (C).

[4] Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India[J]. Parul Arora,Himanshu Kumar,Bijaya Ketan Panigrahi. Chaos, Solitons and Fractals: the interdisciplinary journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena . 2020.

[5] Probabilistic forecasting with temporal convolutional neural network [J] . Yitian Chen,Yanfei Kang,Yixiong Chen,Zizhuo Wang. Neurocomputing . 2020.

[6] Li Jun, XIA Song-zhu, LAN Hai-yan, Li Shou-zheng, Sun Jianguo. Network Intrusion Detection Method Based on GRU-RNN [J]. Journal of Harbin Engineering University, 201,42(06):879-884.

[7] Meng Qiujing, Yang Gang Fault diagnosis method of hydraulic pipeline based on LSTM neural network model [j/ol] Electromechanical engineering: 1-9 [2022-07-17].

[8] Yin Boxin, Yuan Xiaofang, Yang Yuhui, Xie Li Residual life prediction of servo motor rolling bearings based on stacked Gru [j] Machine tools and hydraulics, 2022,50 (12): 153-158.