Autocorrelation Sequence Prediction Model Based On Reference Function Transformation: Taking Epidemic Prediction As An Example

Tingzhen Liu, Tong Zhou, Jin Gao, Wei Li, Yimin Ma
College of Information Science and Engineering, Shenyang University of Technology, Shenyang 110020
E-mail: firstsg@outlook.com

2020 Chinese Automation Congress (CAC) | 978-1-7281-7687-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/CAC51589.2020.9327602

Abstract *Autocorrelation sequence prediction is one of the hotspots in machine learning and statistics. At present, the problem of epidemic prediction is concerned by the whole world in this field. In this paper, a prediction algorithm of autocorrelation sequence based on transformation is proposed. It construct a reference function based on the time series of regions that have experienced the whole process of epidemic situation. The reference function coordinates are transformed with the incomplete observation data of other regions as the supervision. Under the condition that the cost of transformation is kept as small as possible, the transformed function can effectively predict the series values that are not observed in other regions. This algorithm needs a little data and has good training stability. On this basis, we study how to use the information provided by other exogenous variables on the basis of autocorrelation prediction to make the model achieve better results. We use the covid-19 epidemic data set provided by Baidu to test our model, and the results show that it has good fitting metrics. It also has better effect in comparison with LSTM epidemic prediction model baseline.*

Keywords *Autocorrelation Sequence, Time Series Analysis, Neural Network, Coordinate Transformation, Transfer Learning, Machine Learning, Epidemic Prediction*

Introduction

At present, most infectious disease models are variants of differential dynamic models, such as SEIR[1], which is a typical example. Compared with SIR[2] model, it further considers that only a part of the infected population has infectious factors, so that it can model infectious diseases with longer transmission cycle. The latent period model of SEIS[3] is further introduced into the model. By analyzing the corresponding characteristic equations, the local stability of disease-free equilibrium and endemic equilibrium is discussed. The existence of Hopf bifurcation of endemic equilibrium is proved. However, such models do not make good use of known time series information. Considering the use of known data, the work of reference[4] uses the traditional AR family time series prediction method to predict the epidemic situation. But this method requires the sequence to be stationary and linear autocorrelation. The work of reference[5] based on recurrent neural network, when a variety of time series features are introduced, it has great instability[6], and it is difficult to interpret the results effectively. We study an autocorrelation sequence prediction model based on transformation transfer, which has the characteristics similar to one shot learning[7], which can efficiently utilize known complete sequences and has interpretable training stability. When a variety of temporal features are introduced, the non-principal element feature can be used for classification, and the matching reference function can be found for transfer modeling. By this method, the complex multi feature time series prediction problem can be decomposed and transformed effectively.

1 Univariate model

This paper's basic assumption is that the whole process of epidemic situation in different regions is similar, but the length and height of curve is different. Therefore, the curves of other regions can be obtained

by mapping coordinates and scaling from known region curve.

Formal representation, set known complete process curve $f(x)$ (Hereinafter referred to as Reference Function). By introducing the coordinate transformation function g and scale transformation function h, the new curve $h(f(g(x^{'})))$ is obtained. Where $x^{'}$ is the independent variable of the new region (in our case, time). At present, the reference function is known. It is necessary to calculate the parameters of g and h, so that the reference function can fit the observed value of the region with the least transformation cost. In other words, $g(x)$ is equivalent to linear transformation matrix A and translation vector B, $h(x)$ is equivalent to scale transformation vector C. In this way, the transformation process is shown as follows:

$$F^{'}(X) = C * F(AX + B) \quad (1)$$

Where $F(X)$ is the vector form of reference function. We hope to make it fit the known curve of the region to be calculated at the least cost. It is in the following form:

$$\min_{A,B,C} \| F^{'}_{true}(X_{observed}) - F^{'}(X_{observed}) \|_2 \quad (2)$$

$$\min_{A,B,C} \| A \| + \| B \| + \| C \| \quad (3)$$

The constraint of the above formula makes the transformation fit the observation data maximum, and the constraint of the under formula makes the transformation cost as small as possible. After the optimization, the curve value outside the sample can also be estimated through the prior shape provided by the reference function while the obtained curve fits the known curve in the current region.

In order to make the transformation complete with the least cost, based on the viewpoint of reference[8], we choose to use the back propagation algorithm to optimize. To do so, we must get the differentiable representation of f, g and h. $f(x)$ can be obtained by any function fitting. Cubic spline interpolation [9] is our recommended method. g and h need to be adjusted in the process of backpropagation, so here we use two simple neural networks to represent.

The schematic diagram of the model is as follows:

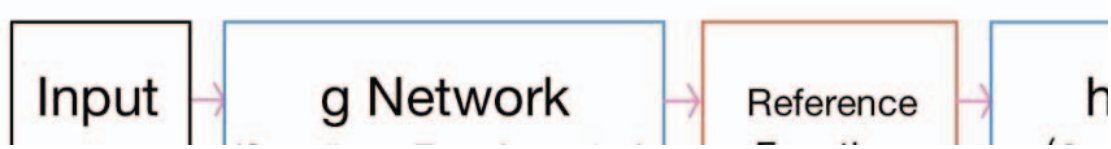

Figure 1. Schematic diagram of Univariate model

In order to make the neural network work better, this paper normalize the data when inputting the sequence into g network and inputting the result of g network into reference function [10]. It should be noted that g and h function prototypes. h represents the scale transformation of reference function, so we need to control its expression ability. We use a simple single linear layer. If the design of h is too complex, the back-propagation process will directly fit the shape of the current region through h, rather than trying to transform the reference function (which is more costly than directly adjusting h [8]). The prototype of g depends on the desired form of coordinate transformation. In our model, g is a simple one-dimensional function of "time-number". From the angle of x-axis, we only want it to shift and zoom left and right, that is, one-dimensional affine transformation. Therefore, only a few linear like layer combinations with bias are tried in the design[11]. Fortunately, even if g is relatively deep, it will not directly affect the results as much as h. Because g only determines the independent variable of f, the final curve value is still calculated by f.

2 Multivariate model

Above this paper analyzed the case of functions of one variable. However, in some infectious disease prediction tasks, there are many features that can be used. How to introduce more features based on the above model is worth considering. This problem can be divided into two cases: one is that the independent variable is a high-dimensional

vector, that is, the reference function is a multivariate function; the other is that the independent variable is composed of a series of time series variables.

2.1 Prediction for multivariate functions

In this case, the above model can still be used, but the difference is that the dimension of the tensor X is changed. A typical example is to study the relationship between x,y coordinates and function values (e.g. the number of infected people varies with regional coordinates), and it is expected to predict the function values of x,y intervals outside the sample. To solve this problem, we focus on the whole 3D graphics described by the reference function $f(x,y)$ (As shown in Figure 2). The transformation form is not different from the previous case of function of one variable. Therefore, the model can be used directly.

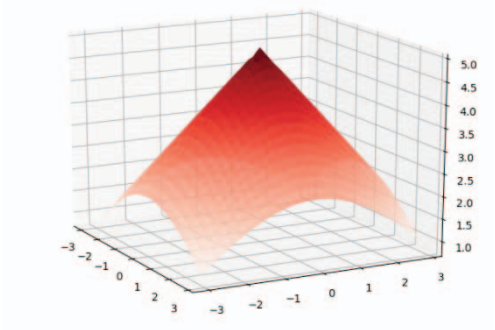

Figure 2. The model can be used when we focus on all the points in it

2.2 Prediction based on multiple time series variables

A typical example of this situation is that the data provider has counted the medical resources m and the number of infected y in each day for several months. They want to predict the number of infections that will follow. In this problem, medical resources, as a time series variable, depend on time. We only care about the predicted value when the value of m is correct. The formula is as follows:

$$y = f(x)|_{m=m_{right}} \quad (4)$$

Where x is the time. m_{right} is the correct value of medical resources, which is given by other prediction models.

The difference from the first case is that we do not want to know the predicted value ($y = f(x,m)$ overall) of medical resources at different values at this time. Instead, it is equivalent to two simultaneous functions:

$$\begin{cases} m = M(x) \\ y = f(x,m) \end{cases} \quad (5)$$

This paper focus on the intersection function of one variable obtained by the above formula. $M(x)$ is the prediction model of medical resources. As shown in the figure below:

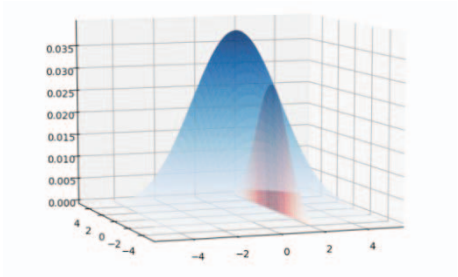

Figure 3. Only the intersection of the red and blue surfaces is focused

3 Experiments

We used the COVID-19 epidemic data set provided by Baidu to test the model. The neural network structure of G and H and the training parameters of the whole model are as follows.

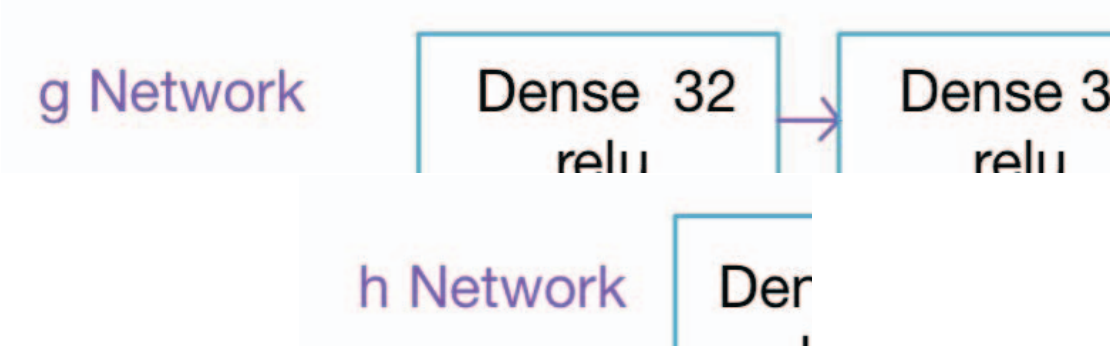

Figure4 . Neural network structure

Table 1. Training parameters

Parameters	Value
Batchsize	45
Epoch	3000
Optimizer	Adam
Loss function	MSE
Evaluation function	RMSLE

We analyzed the data of many epidemic areas and found some representative areas to fit different benchmark functions. We design a classifier, which can find the known regions similar to the regions to be predicted. Then we will use the benchmark function fitted by the regional data to carry out transform and predict the development of the epidemic situation. The whole model is shown below.

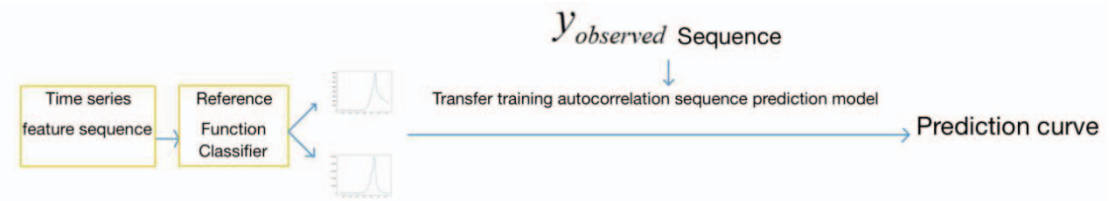

Figure 5. Overall structure diagram of the model

We use a trick which can effectively enhance the model fitting ability: at the junction of reference function and h, the function value of reference function is scaled by the ratio of the maximum value of reference function and the maximum value of $y_{observed}$ k times. The formula is as follows:

$$ref'(x) = \frac{k\max(y_{observed})}{\max(ref(x))} ref(x) \quad (6)$$

$k\max(y_{observed})$ is an estimation of the maximum value of the curve in this region. Use this estimation to preliminarily scale the calculated value of reference function. In this way, the pressure on the scaling function is reduced, and the model converges to the correct direction more quickly.

The reference function does not change during transform. On this basis, g network and h network are trained. Considering the characteristics of the prediction object, we use RMSLE as the evaluation function

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(y_i+1)-\log(y_{true\_i}+1))^2} \quad (7)$$

The loss function has the property that the penalty for under prediction is greater than that for over prediction. In practice, the cost of over estimation is often less than that of under estimation[12].

3.1 Comparative study with LSTM

This paper used LSTM model as the baseline for comparative study. The network structure and training parameters with better indexes are obtained as follows:

Table 2. Network Structure of LSTM

INPUT		
LSTM	128	RELU
LSTM	64	RELU
DENSE	64	RELU
DENSE	64	RELU
DENSE	32	RELU
OUTPUT		

Table 3.Training parameters of LSTM

Parameters	Value
Batchsize	128
Epoch	6000
Optimizer	Adam
Loss function	MSE
Evaluation function	RMSLE
Loss function	MAE
Evaluation function	RMSLE

The fitting results of two representative regions are shown in the figure below:

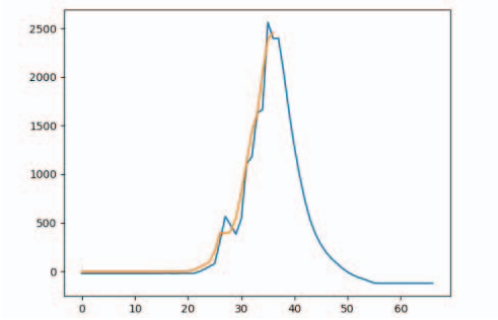

Figure 6.Prediction of LSTM model in areas with severe epidemic situation

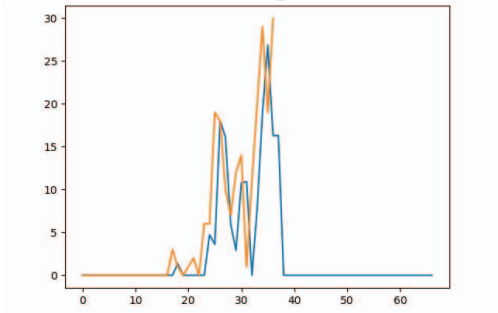

Figure 7. Prediction of LSTM model in low risk areas

The abscissa represents the number of days, and the ordinate represents the number of new infected persons on that day. The orange curve is the known data, while the blue curve is the data predicted by the model.

It can be seen that although he effectively captures the transformation trend, the predicted value drops too fast.

The prediction model based on transformation is used to predict the two regions respectively. The results are as follows:

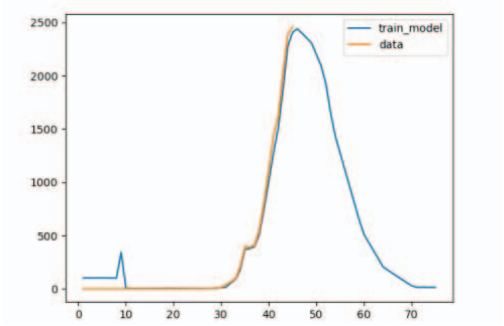

Figure 8. The prediction of our model in areas with severe epidemic situation

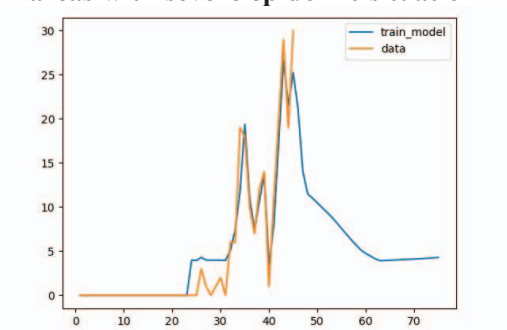

Figure 9. Prediction of our model in low risk areas

As can be seen from the above figure, the model based on transformation not only fits better, but also has a slow downward trend, which is more in line with the real situation. The advantage is also reflected in the evaluation results. The average rmsle loss of all regions is 3.08 when using the LSTM model, and 1.82 is using the transformation based model.

Table4. Results of comparative experiments

Data set	Experimental method	Deep learning framework	Loss Evaluation Loss
Baidu COVID-19 Epidemic data set	Multivariate transformation model	Tensorflow	1.82
Baidu COVID-19 Epidemic data set	LSTM	Tensorflow	3.08

Conclusion

In this paper, an autocorrelation sequence prediction algorithm is studied. In the task of epidemic prediction, the performance is better than the model using LSTM. The model has been extended to apply to prediction task for multivariate function and prediction based on multiple time series variables. The model has the advantages of better use of time series data, decomposition and transformation of complex problems, training stability of theoretical interpretation and efficient use of known information. It can play a certain role in the application research of sequence prediction in the future.

Reference

[1] Biswas M H A , Paiva L T , Pinho M D . A SEIR model for control of infectious diseases with constraints[J]. Mathematical Biosciences & Engineering, 2014, 11(4):761-784.

[2] Tissenbaum HA, Guarente L. Increased dosage of a sir-2 gene extends lifespan in Caenorhabditis elegans.[J]. nature, 2001, 410(6825):227-230.
[3] Xu R , Zhang S , Zhang F . Global dynamics of a delayed SEIS infectious disease model with logistic growth and saturation incidence[J]. Mathematical Methods in the Applied ences, 2016, 39(12):3294-3308.
[4] Yi YF. Epidemic prediction of infectious diseases based on time series model[D]. Changchun University of Technology, 2016.
[5] Sangwon C , Sungjun K , Donghyun L . Predicting Infectious Disease Using Deep Learning and Big Data[J]. International Journal of Environmental Research & Public Health, 2018, 15(8):1596.
[6] Ghazi M M , Nielsen M, Pai A, et al. On the Initialization of Long Short-Term Memory Networks [EB / OL]. Arxiv, 2019.
[7] Vinyals O , Blundell C , Lillicrap T , et al. Matching Networks for One Shot Learning[J]. 2016.
[8] Andreas S.Weigend,D. E. Rummelhart，BernardoA. Huberman. Back-propagation, weight elimination and time series prediction[J]. Connectionist Models Proceedings of the 1990 Summer School, 1991, 105-116.
[9] Boor de,C.Bi-cubic spline interpolation [M]// Interpolating cubic splines /. Birkh user, 1962.
[10] Xiao-Tong L . Study on Data Normalization in BP Neural Network [J]. Mechanical Engineering & Automation, 2010.
[11] Maa C Y , Shanblatt M A . Linear and quadratic programming neural network analysis[J]. IEEE Transactions on Neural Networks, 1992, 3(4): P. 580 - 594.
[12] Wang H , Cui Z , Chen Y , et al. Predicting Hospital Readmission via Cost-Sensitive Deep Learning[J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2018, 15 (6) : 1968 - 1978.