2020 5th International Conference on Information Science, Computer Technology and Transportation (ISCTT) | 978-1-7281-8575-0/20/$31.00 ©2020 IEEE | DOI: 10.1109/ISCTT51595.2020.00072

Analysis and Prediction of COVID-19 in Xinjiang based on Machine Learning

Yunxiang Liu
Shanghai Institute of Technology
SIT
Shanghai, China
yxliu@sit.edu.cn

Yan Xiao
Shanghai Institute of Technology
SIT
Shanghai, China
Corresponding author: 571128791@qq.com

***Abstract*—Covid-19 has taken the world by storm, dramatically affecting the lives of people around the world. China is a major country in the fight against the epidemic. It has provided the world with a wealth of valuable experience in the prevention and treatment of COVID-19.Based on the data released by Xinjiang Health Commission, this study used mathematical modeling method to reasonably predict and analyze the trend of the number of coVID-19 confirmed in the recent outbreak in Xinjiang through machine learning polynomial regression under limited data conditions, aiming at the coVID-19 outbreak in Xinjiang in July.**

Keywords: COVID-19, estimates of the number of confirmed cases, Machine learning, Polynomial regression.

I. INTRODUCTION

In the winter of 2019, cases of viral pneumonia were found in China. On January 12, 2020, the WORLD Health Organization officially named it 2019-NCOV [1]. Soon, the epidemic broke out rapidly in China and confirmed cases were found all over the country. As of September 8, 2020, 90,580 patients have been confirmed in China [2].During the spread of the epidemic, it is of great significance to use effective data for analysis and modeling and make reasonable prediction through machine learning for the analysis and prediction of the entire epidemic [3].In this epidemic, many scholars used different algorithms and achieved good results, making great contributions to the country's fight against the epidemic. Curve fitting, neural network, linear regression and other algorithms have been applied to the prediction and analysis of epidemic situation. In the process of analysis and application, different algorithms will produce different prediction effects, and the data sources and estimated parameters used by the analysis users will affect the accuracy of the prediction results. Better prediction effect can be achieved by building a reasonable model, considering the actual situation as much as possible and setting relatively reasonable parameters.

II. EASE OF USE

A. Experiment Preparation

A large number of studies by many scholars have shown that the incubation period of COVID-19 is about 14 days [4], so the incubation period in this study is considered as 15. Ruguo Fan, Jie Cai and other scholars have pointed out that different incubation periods will make the peak of coVID-19 arrive earlier or later [5].

The SEIR model, which is the most widely used dynamic model of infectious diseases, was used in the experiment. Since the outbreak in Xinjiang received great attention from the state from the very beginning, relevant personnel were quickly quarantined and given timely treatment. Therefore, the model becomes more complex and there are more uncertain factors. We adopt polynomial regression fitting data, use mean square error as the loss function, and use gradient descent algorithm to optimize and obtain the fitting curve.

Finally, the fitting curve is used to predict Due to the strong intervention of the state, relevant personnel were timely isolated and treated, which largely limited the spread of COVID-19 in Xinjiang. Therefore, this experiment ignored the follow-up work and only made short-term prediction based on the strong actual prevention and control situation.

B. Seir Mathematics Model

In 1906, Hamer used a discrete model to study the recurrent epidemic of measles [6]. The SEIR model is a basic mathematical model for infectious diseases. In this epidemic, a large number of scholars use the SEIR model to preliminarily predict the basic reproduction number of the epidemic [7]. SEIR model divides the research objects into S, E, I and R. Where S is the susceptible person, E is the exposed person, I is the infected person and R is the recovered person Susceptible persons are those who do not have the disease but are vulnerable to infection after contact with an infected person because of their lack of immunity. An exposed person is someone who has been in contact with an infected person but is temporarily unable to transmit the infection to others, which is a very important part of an infectious disease with a long incubation period. An infected person is a person who is already infected with an infectious disease. An infected person is contagious and can transmit the infection to an S person, making it either E or I.R is a convalescent, a person who has become immune after a period of isolation or cure. If the immune period is limited, class R members can be converted back to Class S. The SEIR model takes into account the exposure, the exposure An infected person who is temporarily incapable of transmission. In this epidemic, a large number of studies have shown that novel Coronavirus has a long incubation period. In addition, with the outbreak of WuHan will be coronavirus propagation is different, the XinJiang outbreak of COVID - July 19, starting from the outbreak, it attaches great importance to by the countries, the state shall adopt the strong intervention, and taking strong measures, with the support of a large number of manpower

DOI 10.1109/ISCTT51595.2020.00072

material resources and the prevention and control experience, with the help of effectively prevent the secondary infection and outbreaks. Through the comparison of data, it can be found that the impact of government intervention measures on the scale of the epidemic is crucial. Isolation is an effective means to prevent the development of the epidemic. With the intervention of the government, the scale of the epidemic has been well controlled. Of course, this model does not take further into account the further prevention and control measures taken by the government and society, such as the closed management of cities, the research and development of new drugs, etc., so the actual number of people may be lower than that predicted by this model. S(t), E(t), I(t) and R(t) are denoted as susceptible, exposed, infected and rehabilitated persons at time T respectively. According to the conservation relation, S(t) +E(t)+I(t)+R(t) ≡ N, where N is the number of individuals of the population.

The change rate of the number of people in the infected state is shown in formula (1).

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = -\frac{\beta S(t)I(t)}{N} \tag{1}$$

The change rate of the number of people in the exposed state is shown in formula (2).

$$\frac{dE(t)}{dt} = -\frac{\beta S(t)I(t)}{N} - \sigma E(t) \tag{2}$$

The change rate of the number of people in the infect state is shown in formula (3).

$$\frac{dI(t)}{dt} = \sigma E(t) - \gamma I(t) \tag{3}$$

The change rate of the number of people in the rehabilitation state is shown in formula (4).

$$\frac{dI(t)}{dt} = \sigma E(t) - \gamma I(t) \tag{4}$$

The following model is established as formula (5).

$$S_t = S_{t-1} - \frac{r_1\beta_1 S_{t-1}T_{t-1}}{N} - \frac{r_2\beta_2 E_{t-1}S_{t-1}}{N} \tag{5}$$

C. Experimental Design

1) Machine learning

Since this century, artificial intelligence technology breakthrough, machine learning is used in the new field, since the outbreak of an epidemic situation, machine learning technology in China, a national outbreak of fully reflected, such as Yang [8] reported based on short - and long-term memory (long - short - term - the memory, LSTM) recursive neural network forecast model, time model using the data of SARS in 2003 has carried on the AI training algorithm, the model to predict the outbreak will peak in late February.Hu et al. used modified Autoencoders ARTIFICIAL intelligence method to predict the number of newly confirmed cases and the cumulative number of cases in more than 100 countries in real time to provide decision support for the prevention and control process [9].Because machine learning algorithms can be learned and trained based on data from different stages of epidemic development, more accurate prediction models and results can be obtained.

This experiment adopts the machine learning nonlinear regression algorithm, which is a supervised learning algorithm. Because this outbreak is under the strong intervention of the state, and China has gained experience in epidemic prevention and control. This time, coVID-19 transmission was different and more difficult to determine, so polynomial regression was used for fitting.

2) Data selection

In this experiment, the official data provided by the national health commission were used to select the data from July 16 solstice 28, and the later data were predicted.

The data is shown in the figure 1.

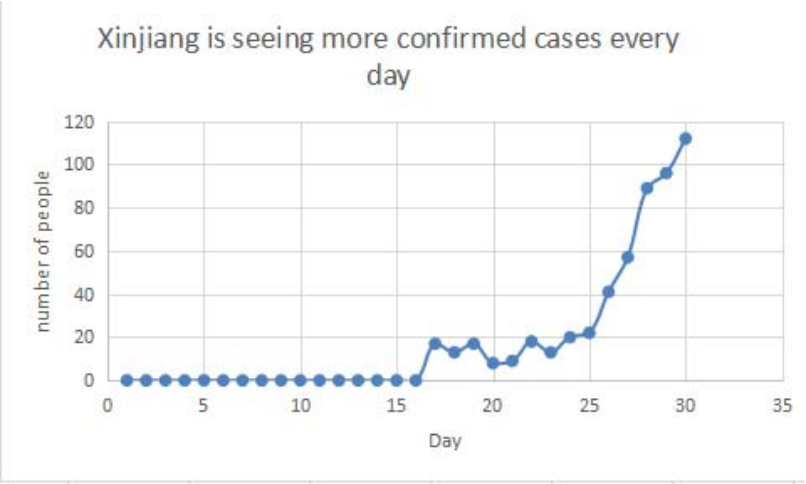

Figure 1. Xinjiang is seeing more confirmed cases every day

3) Loss function

x_i is used to represent the number of days since the outbreak began. It is crucial to select a reasonable loss function when fitting the data. In this experiment, considering the actual situation of the data, we decided to use the mean square error (MSE) to measure the mean value of the square distance

between the predicted value of the model and the true value of the sample y_i. Where, $J(\theta)$ is the loss function and Yi is the number of confirmed cases per day.

The formula of loss function is as formula (6).

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 \quad (6)$$

4) Gradient descent

Objective function is convex. The stagnation point of the objective function is its overall optimal solution. The basic principle of gradient descent method is to iterate continuously along the direction of negative gradient and set the termination error before the optimization begins. If the calculation error is less than the end stop iteration, or to continue the iteration, solving continuously, until they get to minimize the loss function, in some cases, the global optimal solution is often not found, the iterative process are likely to fall into the objective function of the local extremum points, therefore, the accuracy of the forecast model parameter may produce certain effect. First, we randomly initialize theta and then iterate along the negative gradient, making the updated theta smaller $J(\theta)$

The gradient descent formula is shown in formula (7)

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta) \quad (7)$$

5) Polynomial regression

Polynomial regression is a kind of nonlinear regression. When the problem is not linear and the function is very complex, polynomial regression is often very effective.

The method of polynomial regression analysis between a dependent variable and one or more independent variables is called polynomial regression. Polynomial regression model is a kind of linear regression model. The addition of higher powers of features to polynomial regression also increases the degree of freedom of the model to capture nonlinear changes in the data. Adding higher-order terms also increases the complexity of the model. Combined with the SEIR model and the actual complexity of this outbreak, we can see that the growth trend is still a curve, but it is more complicated than before. Combined with the constructed model, this experiment decided to use the fourth degree polynomial for fitting. We use polynomial regression to fit the data we collect.

III. Results and Discussion

A. Experimental procedure

The first step is to collate the data we have obtained and process it programmatically in Python. In the second step, we input the data into the polynomial for fitting, and continue to optimize according to the size of the result of the loss function until we find the most suitable fitting curve. The third step is to make a prediction for the next few days based on the obtained functions and compare the predicted results with the actual results.

The fitting effect is shown in the figure2

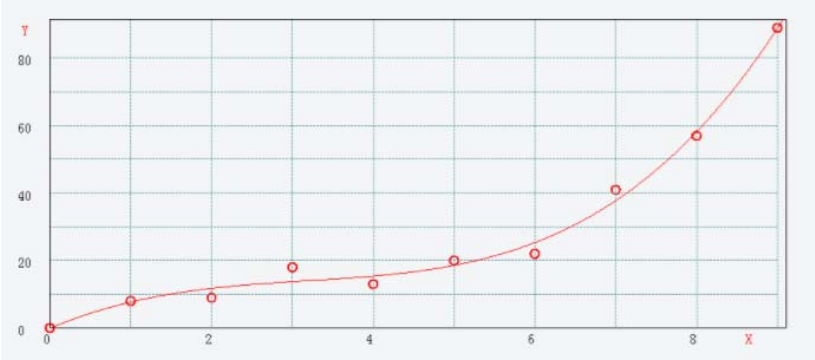

Figure 2. Fit function diagram

The prediction effect is shown in the figure 3.

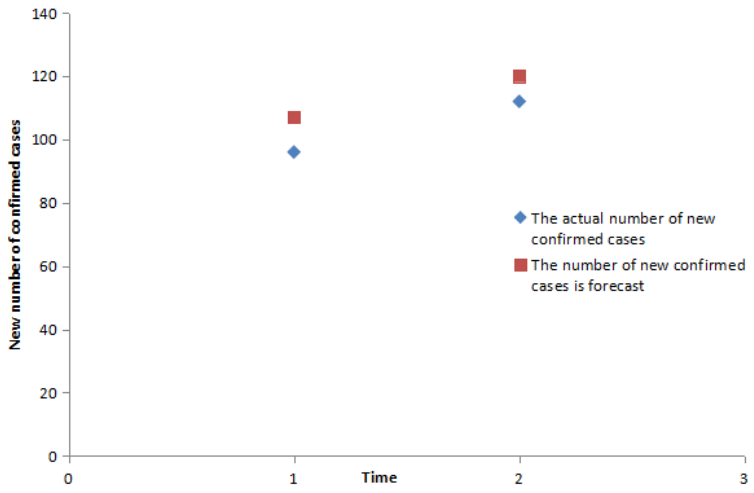

Figure 3. Prediction renderings

B. *Interpretation of result*

From the experimental results, it can be seen that the fitting effect is more accurate in predicting the number of new people within two days, and can make a good prediction for the development of the epidemic. The experiment also shows that the nonlinear function can be fitted well by using polynomial regression under complex and variable conditions.

IV. CONCLUSION

In this study, the transmission model of COVID-19 was modeled, and the regression curve was obtained by multinomial regression according to the daily number of coVID-19 diagnosed in Xinjiang. We can see that in the short-term prediction, the use of polynomial regression can better predict the number of new confirmed cases per day. At the same time, we can also see that the scale of the epidemic has been well controlled with the strong intervention of the state and people's enhanced awareness of self-protection. Compared with foreign outbreaks in the same period, the level of national awareness of prevention and the timely isolation and adoption of strong control measures will have a profound impact on the epidemic [10]. The success of Epidemic prevention and control

In China also fully demonstrates that novel Coronavirus can be prevented and controlled. Unlike the epidemic in Wuhan, the epidemic has been well controlled under the strong national policy, which fully demonstrates the importance of timely isolation. A large number of experiments have shown that timely isolation and identification of related close contacts after novel Coronavirus infection can effectively control the epidemic. As the epidemic is not over yet, with the deepening of the study, the data on the epidemic will be more accurate and the prediction will be more accurate. With the emergence of other asymptomatic infected persons, the study model needs to be further optimized.

REFERENCES

[1] Hsieh Y H, Chen C W S, Hsu S B . SARS outbreak, Taiwan, 2003[J]. Emerging infectious diseases, 2004, 10 (02) :201 .

[2] National Health Commission, PRC.Novel Coronavirus infection pneumonia.outbreak.update.http://wwwnhcgovcn/xcs/yqtb/list_gzbd.shtml.

[3] Effective control measures under the novel Coronavirus fashion trend simulation [J]. Bai Ruhai, Dong Wan Yue, SHI Ying, FENG Aozi, LI Li, XU Ding, LU Jun. Medical knowledge. 2020(02)

[4] National Health Commission, PRC.Novel Coronavirus Infection Prevention and control Protocol ［Z］ .

[5] Hsieh Y H， Chen C W S， Hsu S B. SA R S outbreak， Taiwan， 2003[J] . Emerging infectious diseases， 2004， 10 (02) :201.

[6] Xu Zhijing, Zu Zhenghu, Xu Qing, et al.Advances in dynamic modeling of infectious diseases.Military medicine，2011，35(11) : 828 — 833.

[7] Zhou Tao, Liu Quanhui, Yang Zimo, et al.Preliminary prediction of novel Coronavirus infected pneumonia with the number of regeneration.Chinese Journal of Evidence-based Medicine，2020，20(3) : 359 — 364.

[8] Yang Z，Zeng Z，Wang K，et al. Modified SEI R and AI prediction of the epidemics trend of COVID — 19 in China under public health interventions，J Thorac Dis，2020，12(3) : 165.

[9] Hu，Z.，Ge Q，Li S，et al. Evaluating the effect of public health intervention on the global-wide spread trajectory of Covid — 19. med R xiv， 2020: p. 2020. 03. 11. 20033639.

[10] R emuzzi A，R emuzzi G. COVID19 and Italy: what next? [J] . The Lancet， 2020， 395:12251228.