ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202111_4(11).0078

Comparative Research on Prediction of COVID-19 Based on SIR Model and Machine Learning Algorithm

Li Xiong[1], Peiyang Hu[1, a]

[1]School of management, Shanghai University, Shanghai, 200000, China

[a]shareworld@foxmail.com

Abstract

The outbreak of new coronavirus pneumonia in 2019, which is called Corona Virus Disease 2019 (COVID-19), has been the most serious infectious disease pandemic in the world in 100 years. It is a major public health emergency with the fastest spreading speed, the widest infection range and the most difficult to prevent and control. Currently, the spread of COVID-19 is still global, and the epidemic situation in some countries is still very serious. Many scholars have also started to crawl, summarize, analyze various kinds of historical data emerging on the network, and use various algorithms to design the epidemic prediction model of new coronary pneumonia. In this paper, based on the confirmed, dead and cured cases of COVID-19 in the United States obtained from the Center for Systems Science and Engineering of Johns Hopkins University, SIR models, logistic regression as well as support vector regression algorithms in machine learning are used to simulate and predict the development of the epidemic, and the accuracy of each prediction model is compared. In order to provide more accurate reference for the follow-up epidemic warning and prevention and control.

Keywords

Corona Virus Disease 2019; SIR Model; Logistic Regression; Support Vector Regression; Prediction Comparison.

1. Introduction

Infectious diseases have been a major test for countries around the world for many years. Infectious diseases have risks such as human-to-human transmission. The rapid spread of infectious diseases, the high degree of harm and the high degree of complexity make them the most difficult to deal with in emergency management. In addition, the weak links such as warning, cross-regional prevention and control, and material allocation exposed in the disposal of epidemic situation make the monitoring, early warning and disposal mechanism of public health emergencies a long-standing problem. The death or trauma caused by them far exceeds the sum of war deaths. With the development of society, human communication has become more and more close, so infectious diseases have a faster and wider transmission than ever before. At present, infectious diseases are also the main diseases of morbidity and death. The outbreak of Corona Virus Disease 2019 (hereinafter referred to as "COVID-19") has swept the world in a hurry. This unexpected public health and safety event has caused great harm to our country and even the world. However, it also reminds us that we should study and summarize this event in depth, and use scientific methods to predict the incidence trend of infectious diseases based on a large number of epidemic data and other data that have a serious impact on the incidence and spread of infectious diseases. Currently, the scientific methods include the self-contained epidemic model and the machine learning methods that have received much attention in recent years. With the rapid development of new generation information technology such as big data, artificial intelligence and machine learning, it is urgent and

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202111_4(11).0078

important to make full use of modern and advanced technology to monitor and warn the epidemic situation. Timely and accurate risk monitoring and prevention and control early warning can effectively carry out emergency management, reminding departments to do well in advance of preventive measures, reduce the various economic and social negative effects of large-scale spread of infectious diseases. However, although with the improvement of machine learning theory and the detection of practice, the ability to handle more complex problems may be stronger, has there been enough capacity to completely replace the traditional infectious disease model? This study will focus on the analysis.

2. Relevant Research

As an important gateway to all information about the new coronary pneumonia epidemic, the Internet has become the main "information source" platform for the epidemic. In this epidemic prevention war, can we build an intelligent epidemic identification model by combining various algorithms in the field of machine learning with computer technology, and bring into play the war "epidemic" value of computer technology?

On the basis of SIR model, Cooper (2020) estimated the parameters with investigating the temporal evolution of populations and monitor important parameters of disease transmission in different communities. The simulation and prediction of the development of severe epidemic situation in countries showed that the model was basically reliable for the analysis of the development trend of COVID-19. Scholars such as Milhinhos (2020) chose the most common supervised learning algorithm in machine learning, non-linear regression algorithm. The least squares criterion and gradient descent method were used to perform non-linear regression on the data to find the non-linear relationship between days and the number of patients diagnosed. Then, the trend of the number of patients diagnosed with COVID-19 was predicted by using mathematical modeling. In addition, based on the SEIR (susceptible-latent-communicator-restorer) model, Radulescu (2020) established an integral universal SEIR model to describe the transmission mechanism of COVID-19 in susceptible populations, and combined with the parameters of average latency, average infectious period and proportion of atypical patients of the new coronavirus pneumonia epidemic. To simulate the global situation of new-type coronavirus pneumonia in major epidemic countries.

Indeed, more and more researchers are starting to carry out data-based epidemic prediction studies using various algorithms to supplement the deficiencies of existing prediction systems. In these studies, large data, such as Internet search queries, are being used abroad to monitor the occurrence of infectious diseases. Internet search data can be collected and processed at a near real-time speed. Towers (2015) found that searching data over the Internet can create infectious disease identification models faster than traditional monitoring systems. In addition, some scholars such as Huang (2018) attempted to use the generalized additive model (GAM) to identify and predict hand, foot and mouth disease, which includes the best results obtained by searching for queried data, and to obtain new large data identification and monitoring tools with the advantage of easy access to the incidence of infectious diseases, which can identify infectious disease trends before official organizations.

In addition to Internet search data, big data from social media is also being considered. Researchers such as Tenkanen (2017) found that large data from social media is relatively easy to collect and can be freely used, which means real-time accessibility, real-time continuous creation of data, and rich content. Therefore, microdata can be used to predict infectious diseases like this COVID-19 in addition to predictions in a variety of other scientific fields. Scholars such as Shin (2016) also found that infectious diseases are highly correlated with data from social media and that it is possible to use digital recognition systems to monitor infectious diseases.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202111_4(11).0078

3. Research Methods

3.1. SIR Model

The SIR model, also known as the Compartment Model, is a classic infectious disease model invented in the early part of the last century that gives a rough picture of the process from the onset to the end of an infectious disease. The model is suitable for diseases transmitted through the virus, such as influenza, measles, chickenpox and other types of infectious diseases that are immune to the original virus after recovery.

In the model, the population within the epidemic range of infectious diseases is divided into three categories: Class S, which means susceptible, refers to the people who do not have the disease, but lack immune ability, and are vulnerable to infection after contact with the infected people; Class I, which means infected, refers to the infected person, which can be transmitted to class s members; Class R, which means removal, refers to people who are isolated or have immunity due to recovery. Assuming that the total number of people in the sample area remains unchanged, N(t)=c (c is a constant), in unit time t, the number of susceptible people that a patient can infect is directly proportional to the total number of susceptible people s (T) in this environment, and the proportion coefficient is β. Thus, the number of people infected by all patients at time t is βS(t)I(t). In addition, the disease can be cured or lead to death, resulting in R(t) population. At time t, the number of people removed from the infected person per unit time is directly proportional to the number of patients, and the proportion coefficient is γ, The number of people removed per unit time is γI(t).

So the flowchart for the SIR model is shown in Figure 1 below:

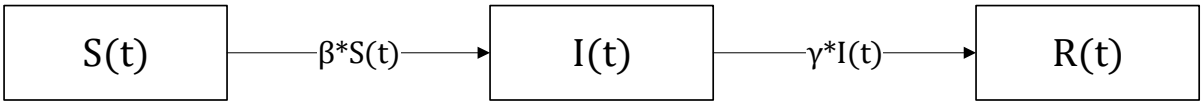

Figure 1. The SIR model flowchart

As can be seen from Figure 1, this model is a one-way model, the number of susceptible people is constantly entering into the number of infected people, and the final number of infected people is also in the number of single-yearities to recover, so the number of susceptibility and the number of infections will eventually drop to 0, while at the same time, all will become the number of people recovered.

The differential equations for the SIR model are:

$$\begin{cases} \frac{dS(t)}{dt} = -\beta S(t)I(t) \\ \frac{dI(t)}{dt} = \beta S(t)I(t) - \gamma I(t) \\ \frac{dR(t)}{dt} = \gamma I(t) \end{cases} \quad (1)$$

3.2. Logistics Regression

Logistic Regression, also known as Logistic Regression Analysis, is a classic classification algorithm in machine learning. Logistic regression is a broad linear regression analysis model, often used in data mining, automatic disease diagnosis, economic prediction and other fields. For example, explore the risk factors that cause disease and predict the probability of disease from being caused by risk factors. Logistic regression has become almost the most commonly used analytical method in epidemiology and medicine, including exploring the risk factors that cause disease and predicting the probability of disease from risk factors. The principle is to calculate the probability of the condition under the logic distribution, and to select the party

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202111_4(11).0078

with the large probability of the condition as the prediction category. Logistic function or Logistic curve is a common S-shaped function, which is widely used in biological reproduction and growth process, population growth process simulation. The function formula is as follows:

$$P(t) = \frac{KP_0e^rt}{K + (P_0e^rt - 1)} \tag{2}$$

Where, P_0 represents the initial value, K represents the final value, R is the growth rate measuring the speed of curve change, and t is time.

3.3. Support Vector Regression

Support Vector Machine (SVM) is a powerful and comprehensive machine learning model capable of performing linear or nonlinear classification, regression, and even outlier detection tasks. SVM is a classic two-classification model, the basic model is defined as the largest interval in the feature space linear classifier, its learning optimization goal is to maximize the interval, so the support vector machine itself can be transformed into a convex secondary planning solution problem. Unlike Logistic regression, SVM does not output the probability of each category.

Support Vector Regression (SVR) is an important application branch of SVM. Short-term predictive modeling using SVR essentially maps a nonlinear map as $\varphi(x)$ a sample to high-dimensional space S for linear regression, which is:

$$f(x) = \omega^T\varphi(x) + b \tag{3}$$

Where weights are and bias vectors, i.e. thresholds.$\omega\varphi$

4. Research Process and Results

In this part of the study, the process in Figure 2 will be followed to reach a completer and more convincing conclusion.

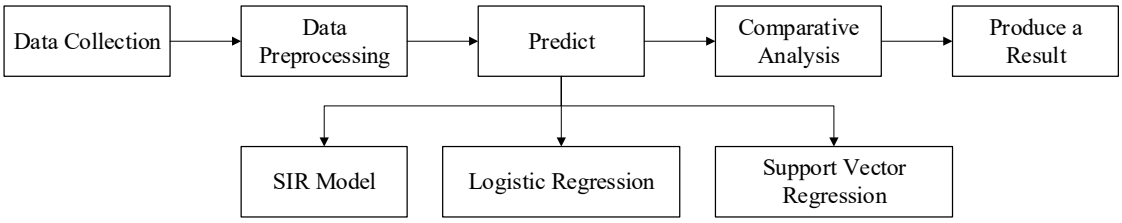

Figure 2. Research Process Flowchart

4.1. Data Source

The research data were obtained from the real-time dynamic web page of Dr. Lilac Garden Lilac's new coronavirus pneumonia outbreak. To ensure the accuracy of the data, the data disclosed by johns Hopkins University's Center for Systems Science and Engineering was also validated before the experiment. After data cleaning, screening and other pre-processing work, the United States from March 1 to May 30, 2020 cumulative diagnosis, new diagnosis, cumulative death and cumulative number of cures, as a training set sample of the study, as shown in Table 1 (https://ncov.dxy.cn/ncovh5/view/pneumonia, https://systems.jhu.edu/)

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202111_4(11).0078

Table 1. Training Set Sample

Date	Cumulative number of confirmed cases	New number of confirmed cases	Cumulative death toll	Cumulative number of cures
March 1, 2020	30	6	1	7
March 2, 2020	53	23	6	7
March 3, 2020	73	20	7	7
March 4, 2020	104	31	11	7
March 5, 2020	174	70	12	7
.....				
May 26, 2020	1689057	18848	99239	384902
May 27, 2020	1707423	18366	100744	391508
May 28, 2020	1730259	22836	101937	399991
May 29, 2020	1754747	24488	103113	406446
May 30, 2020	1778993	24246	104054	416461
May 31, 2020	1799122	20129	104659	444758

Based on the data set, a graph of the development of the number of COVID-19 confirmed cases in the United States from March 1 to May 30, 2020 is mapped, as shown in Figure 3. It is intuitive to see that the new crown epidemic in the United States is developing rapidly. As things stand, although the "New Coronation Pneumonia" epidemic in the United States spread faster, infected, and killed more people than any other country in the world, the United States itself has not really reached its peak infection. This makes predictions of outbreaks in the United States meaningful.

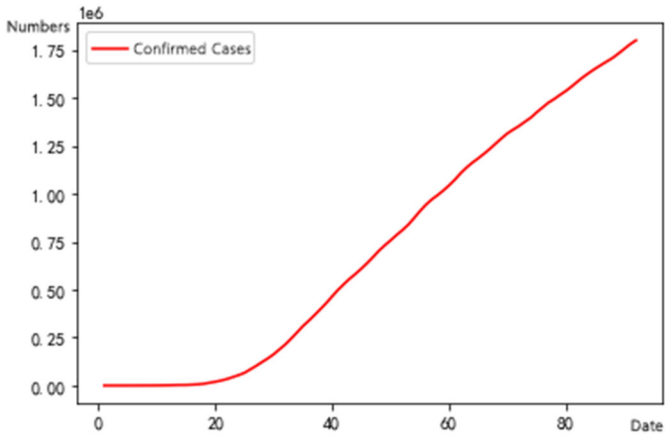

Figure 3. Graph of Data on the Number of Confirmed Cases of COVID-19 In US From March 1 To May 30, 2020

4.2. Use the SIR Model to Make Predictions

The search revealed that the population of the United States is about 330 million, and data sets can be found that the total number of confirmed cases in the United States on March 1st was 30, bringing the cumulative number of cures and deaths to a total of eight. Because in practice, both the cure and the death case represent that the patient will not pass the disease on to others in the future, the combined cure and death toll at the time of the study is the number of recoveries in the model.

Then, the study needs to determine the time t and the two parameters in the differential equation β and γ. As the population of infectious diseases is spreading, the development situation and value of the epidemic situation are closely related. Then, the study will use the

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202111_4(11).0078

population data of the United States as an example to see how different values will change the epidemic situation without external intervention.

In the experiment, four different sets of values will be set up to predict and compare the results:

Table 2. Sets of Different Value

Constituencies	βvalue	γvalue
1	0.125	0.05
2	0.25	0.05
3	0.25	0.1
4	0.5	0.1

Of the four sets of predictions, the first and second sets of values were obtained by searching for research reports on COVID-19. In most reports, estimates for β ranged from 0.125 to 0.25, while estimates for γ were approximately equal to 0.05. This means that the cure period of COVID-19 is about 20 days. In the first group, the β value was set at a low 0.125, and in the second group, the β value was doubled, while the γ value was unchanged. In other words, in the second estimate, the COVID-19 was more infectious than in the first, but the rate of cure did not change, more in line with the current situation in the United States. In the third group, we doubled the rate of recovery while maintaining a β value of 0.25, and patients recovered in half the time. In the fourth group, the infection rate of the COVID-19 was doubled, and the patient was set to recover at half the usual rate.

The experimental results are shown in Figure 4. In these four graphs, the prediction of the epidemic situation in the United States is significantly different with the variation of β and γ values, and the epidemic peak is affected by the β/γ value, that is, the basic regeneration number. The value of the Basic Reproductive Number (R_0) represents the average Number of new infections resulting directly from a typical case in a population of all susceptible individuals. In the 4 groups, R_0 is 2.5, 5, 2.5, 5. According to the research of Steven (2020), the value of the second group is closer to the reality.

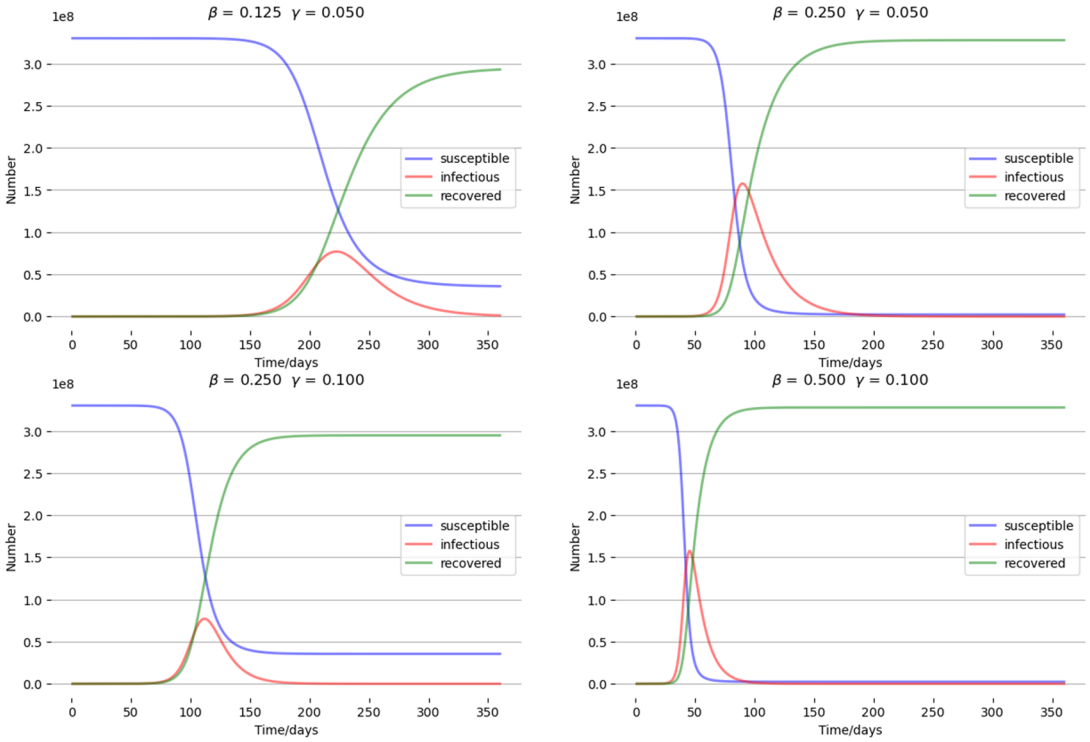

Figure 4. A Result Graph Using SIR Model for Prediction with Different Values

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202111_4(11).0078

In the SIR model prediction, if β=0.25 and γ=0.05 are used, it can be seen from the figure on the upper right of Figure 4 that the EPIDEMIC in the United States will peak approximately 90 days after May 31, and then start to decline, and the number of infected people will reach nearly 150 million at the peak.

4.3. Use Logistic Regression to Make Predictions

Using LR for prediction is essentially fitting a Sigmoid curve using Logistic function expression. According to the analysis of Logistic regression function in 3.2, the initial value P_0, final value K, growth rate r and time T need to be solved by Logistic function. The fitting parameters can be obtained by using the curve_fit function in the optimizer module of Scipy in Python. But we still need to determine the value of K.

As can be seen from Figure 3, the values changed slowly in the early stages of the outbreak, indicating that the U.S. outbreak was only very small at this time and that the government did not pay enough attention to it. After the outbreak phase began, the value increased dramatically, indicating that the epidemic spread rapidly and that the government began to pay more attention to testing. But as of the end of May, the U.S. outbreak showed no signs of an inflection point, so in this study, in order to get more accurate results, the values were repeatedly debugged, and $KKKK = 100000$finally, when the closest to the late April to the end of May diagnosis, as shown in Figure 5.

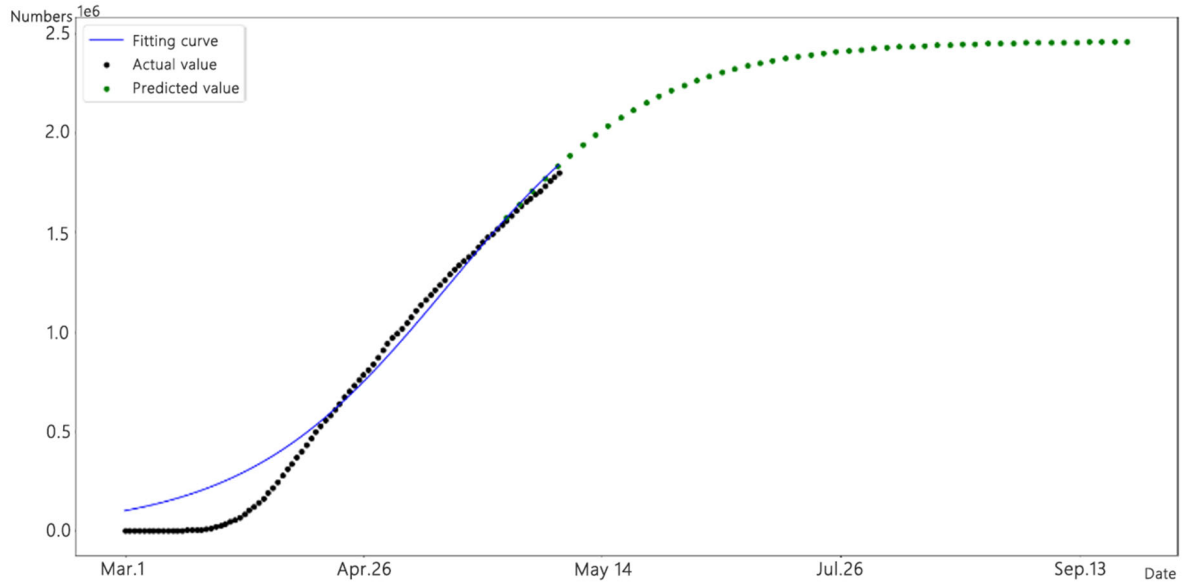

Figure 5. A result Graph Using Logistic Regression for Prediction (K=100000)

In projections using the Logistic Regression model, the graph in Figure 5 shows that the U.S. outbreak peaked around the end of August and then began to decline, with nearly 25 million people infected at the peak.

4.4. Use support Vector Regression to Make Predictions

Support Vector Regression (SVR) is an important application branch that supports Vector Machines (SVMs). But there is a difference. SVM classification is to find a regression plane that allows the support vectors of the two classification sets or all the data to be the furthest away from the classification plane, while SVR regression, on the contrary, is to select a portion of the training data to support vectors more effectively and predict targets through regression analysis based on the values of these training samples. Or, to find a regression plane, so that all the data from a collection is closest to that plane.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202111_4(11).0078

When making predictions with SVR predictions, similar to using Logistic regression, you need to find the optimal parameters first. In this study, the RandomSearchCV function of the Scikit-learn module in Python was used to randomly search for hypers parameter space, use mean square error as an evaluation index, train 30 rounds, and finally establish a support vector regression model. Finally, the optimal combination of parameters is applied to the model to make predictions, and the prediction values are shown in Figure 6. In the figure, the black line is the true value of the original data, and the blue line is the prediction value of SVR, so it can be seen that the change rate of the predicted result is smaller than the actual value.

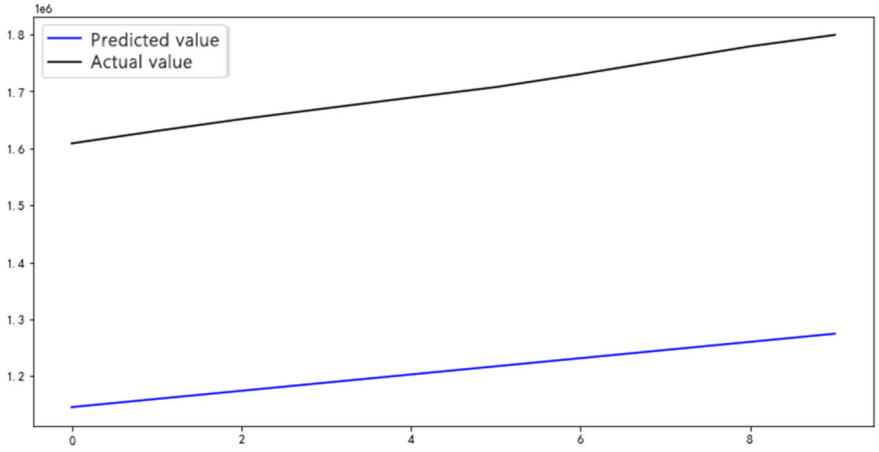

Figure 6. Use SVR to Get the Predicted Values

Finally, the actual values and the results obtained using SVR prediction are plotted to get the results shown in Figure 7, where the SVR prediction is represented directly by the dashed blue line. As can be obtained from Figure 7, a regression line is indeed found, making all current data closest to the plane. However, the forecast results are not ideal, and the resulting forecast value may be significantly low.

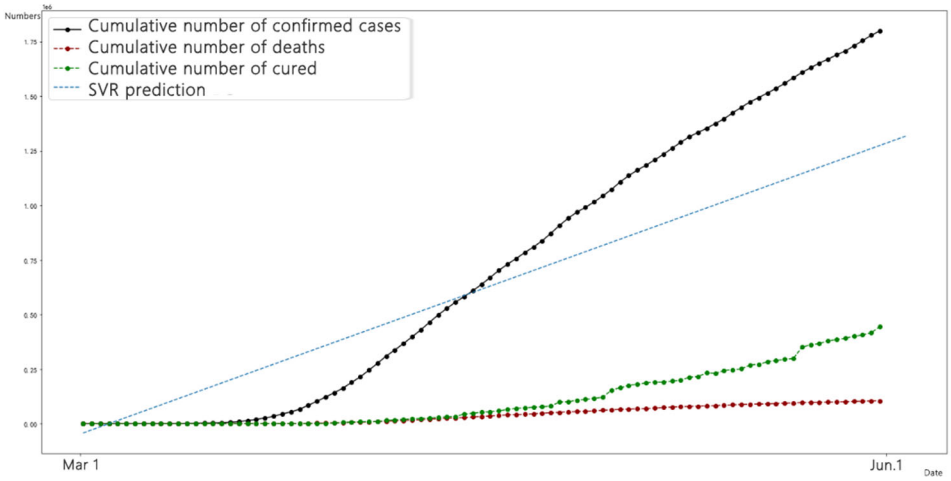

Figure 7. A Result Graph Using SVR for Prediction

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202111_4(11).0078

5. Conclusion and Discussion

5.1. Conclusion

There are many models and algorithms that can be used for prediction analysis. Based on the time series dataset collected, this paper selects the US data with more severe epidemic situation, and uses SIR model, Logistic regression algorithm and SVR algorithm to predict. The results showed that the SIR model and logistic regression algorithm predicted the peak time of the epidemic more closely. However, since SVR only produces a regression line, it is not helpful to predict the peak value. On the prediction of the cumulative number of confirmed cases of the epidemic, the gap between the three appears large. The logistic regression algorithm is in good agreement with the current reality, while the SIR model and the SVR algorithm predict too high and too low data, respectively.

In fact, the SIR model, logistic regression algorithm and SVR algorithm only predict based on the cumulative number of confirmed cases, the cumulative number of deaths, the cumulative number of cured cases and other data. On the one hand, they are more affected by latency, policy control and other factors such as drugs. Therefore, the results of this model alone may be inconsistent with the actual situation. On the other hand, all three models have parameters that need to be manually adjusted, and the setting of parameters may also affect the accuracy of prediction.

Nevertheless, this study is of practical significance by modeling the spread of new coronavirus pneumonia and predicting the time of inflection point according to the relevant data. Either model confirms that there is still a long way to go before the inflection point of the COVID-19 outbreak in the United States, which also warns the public already in the United States should avoid places where crowds gather when they go out and take self-protective measures to pay attention to wearing masks.

5.2. Limitation and Prospect

This paper combines various prediction models to make a comprehensive analysis of the data of the new type of coronavirus pneumonia in the United States, and to obtain more comprehensive results. However, there are some shortcomings. Firstly, the "new coronary pneumonia" virus itself has a certain latency period, and healthy people who have come into contact with the infected person do not immediately get sick, but become carriers of the pathogen. Therefore, the Exposed category should be added to the SEIR model in theory. However, when data was collected, the relevant records of infected persons abroad could not be crawled, so the prediction could not be made. In future studies, further improvements should be made to existing models to find alternative expressions of infected persons in order to better suit the actual situation.

Secondly, in the use of Logistic regression model to predict the development of the epidemic, in order to better fit the second half of the model, ignored the development of the pre-epidemic. Under the influence of the outbreak period and the intervention of the U.S. government, the development process of the epidemic may be different from the front and back, so it may be possible to model and analyze the situation of the U.S. epidemic before the outbreak, select the appropriate model, and then optimize the model equation and adjust the parameters to achieve a better fit effect, and more accurately describe the whole process of the development of the U.S. epidemic.

Acknowledgments

We acknowledge the supports from the National Social Science Foundation of China (No. 21ZDA105).

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202111_4(11).0078

References

[1] Cooper, I. , Mondal, A. , & Antonopoulos, C. G. . (2020). A sir model assumption for the spread of covid-19 in different communities. Chaos Solitons & Fractals, 139, 110057.

[2] Milhinhos, A. , & Costa, P. M. . (2020). On the progression of covid19 in portugal: a comparative analysis of active cases using non-linear regression. Frontiers in Public Health, 8, 495.

[3] Radulescu, A. , & Cavanagh, K. . (2020). Management strategies in a seir model of covid 19 community spread. Scientific Reports.

[4] Sherry, T. , Shehzad, A. , Gilbert, B. , Nadya, B. , Shala, B. , & Baltazar, E. , et al. (2015). Mass media and the contagion of fear: the case of ebola in america. PLoS ONE, 10(6), e0129179.

[5] Da-Cang, Huang, Jin-Feng, & Wang. (2018). Monitoring hand, foot and mouth disease by combining search engine query data and meteorological factors. Science of The Total Environment, 612, 1293-1299.

[6] Tenkanen, H. , Minin, E. D. , Heikinheimo, V. , Hausmann, A. , & Toivonen, T. . (2017). Instagram, flickr, or twitter: assessing the usability of social media data for visitor monitoring in protected areas. Scientific Reports, 7(1).

[7] Shin, S. Y. , Seo, D. W. , An, J. , Kwak, H. , Kim, S. H. , & Gwack, J. , et al. (2016). High correlation of middle east respiratory syndrome spread with google search and twitter trends in korea. Scientific Reports, 6(1), 32920.

[8] Bjornstad, O. N. , Finkenstadt, B. F. , & Grenfell, B. T. . (2002). Dynamics of measles epidemics: estimating scaling of transmission rates using a time series sir model. Ecological Monographs, 72(2), 169-184.

[9] Scrucca, L. . (2015). Model-based sir for dimension reduction. Computational Statistics & Data Analysis, 55(11), 3010-3026.

[10] Yiu, T. W. , Cheung, S. O. , & Chow, P. T. . (2008). Logistic regression modeling of construction negotiation outcomes. IEEE Transactions on Engineering Management, 55(3), 468-478.

[11] Muller, Clemma, J. , MacLehose, & Richard, F. . (2014). Estimating predicted probabilities from logistic regression: different methods correspond to different target populations. International Journal of Epidemiology.

[12] Yu, P. S. , Chen, S. T. , & Chang, I. F. . (2006). Support vector regression for real-time flood stage forecasting. Journal of Hydrology, 328(3-4), 704-716.

[13] Basak, D. , Srimanta, P. , & Patranbis, D. C. . (2007). Support vector regression. neural information processing letters & reviews.

[14] Sanche, S. , YT Lin, Xu, C. , Romero-Severson, E. , Hengartner, N. , & Ke, R. . (2020). High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. Emerging Infectious Diseases, 26(7).