Proceedings of the 39th Chinese Control Conference
July 27-29, 2020, Shenyang, China

Prediction of COVID-19 spread via LSTM and the deterministic SEIR model

Yifan Yang, Wenwu Yu, Duxin Chen
School of Mathematics, Southeast University, Nanjing 210096, P. R. China
E-mail: chendx@seu.edu.cn

Abstract: Due to the outbreak of COVID-19, China and most countries in the world have been seriously affected and tens of thousands of people have lost their lives, it is urgent to study the transmission characteristics and trends of the virus. In this study, we adopt the Long Short Term Memory algorithm at first to predict the infected population in China. However, it does not explain the dynamics of diffusion process, and the long-term prediction error is too large. Therefore, the widely-accepted SEIR model is introduced to capture the spread process of COVID-19. By using a sliding window method, we suggest that the parameter estimation and the prediction of the infected populations are well performed. This may provide some insights for epidemiological studies and understanding of the spread of the current COVID-19.

Key Words: Epidemiological model, LSTM, Sliding window method, COVID-19

1 Introduction

As human beings are facing long-term and serious threats from various diseases, such as H1N1, H7N9 and the current COVID-19, the spread of the epidemic has seriously affected our lives and national economy[1]. The spread of the epidemic is full of accidental factors, which is obviously a complex dynamic phenomenon[2]. The small-world network feature of our living society make the spread radius of diseases or rumors larger and larger, which makes it easier to spread diseases[3]. Thus, it is of great practical significance and urgency to study the mechanism of epidemic spreading, which would become the basis for intervention and an important way to suppress the epidemic.

In recent years, in order to explore the spread of virus in different population structures, many epidemiological models have been proposed, such as the compartment model [4] for small-scale, well mixed population and the network epidemiological model [5] for individuals with complex contact relationship in the regional population. At present, the most widely used model for epidemic spread in large-scale spatial regions is the metapopulation model. Previous studies [6, 7] used the data of global aviation networks and mobility network of mobile phone users to analyze the spread of SARS and H1N1 in global urban population. However, in practical application, it is usually impossible to obtain the detailed flow data of all cities, so most of the previous studies can only use the coarse-grained flow data to establish epidemic transmission networks between cities.

Meanwhile, with the rapid development of AI techniques and network science tools, it is feasible to reveal the evolutionary rules of such complex systems[8]. For instance, Fraser *e*t al. used limited data to determine the pandemic potential of H1N1 [9], and conducted an early assessment of the transmissibility and severity of the outbreak of international spread. Wang *e*t al. proposed a network inference model [10], which can reduce the individual based network into a sub-population network without losing information. In addition, in order to reveal the dynamic mechanisms of disease transmission, many recent studies use network analysis strategies to predict the spread of epidemics in social systems [11, 12].

As we know, COVID-19 spreads rapidly to other cities and provinces in China and international countries due to the large-scale population flow in the early days. In this paper, we aim to study the dynamics and prediction of the trend of COVID-19, and hope to provide some knowledge of this disaster. We first test and verify that the current prediction method based on neural network can predict the evolution trend of different types of population. Then, we introduce the widely accepted SEIR model and identify the parameters to fit and predict the future trend. We suggest that by using the sliding window method based on the proposed SEIR model, we can predict the inflection and endpoint of COVID-19.

Throughout this paper, considering that the actual spread of COVID-19 is determined by many factors, we use the following assumptions to simplify the study, but without loss of generality:

(1) It is assumed that virus transmission occurs in a closed environment, independent of natural birth rate and natural mortality.
(2) The data of the confirmed, cured and dead cases are basically accurate.
(3) The patient is not infectious during latent period and there is no super-spreader.
(4) The cured will be able to produce antibodies to prevent them from reinfecting.

2 The Single-region SEIR Model

In this section, we introduce the classical SEIR model [13], which will be used to describe the recent outbreak of COVID-19 in China. We use $S(t)$ to represent the susceptible population, $E(t)$ to represent the exposed population, $I(t)$ to represent the infected population, $R(t)$ to be the removed population, including those who are cured or dead due to the COVID-19, and the total population is N. Let the time-dependent variables be simplified without index t, and the dynamics of the model is described below:

This work was supported by the Fundamental Research Funds for the Central Universities (Grant No. 2242019K40111), the National Natural Science Foundation of China (Grants Nos. 61903079 and 61673107), and the Jiangsu Provincial Key Laboratory of Networked Collective Intelligence (Grant No. BM2017002).

$$\begin{cases} \dfrac{dS}{dt} = -\dfrac{\beta SI}{N}, \\ \dfrac{dE}{dt} = \dfrac{\beta SI}{N} - \sigma E, \\ \dfrac{dI}{dt} = \sigma E - \gamma I, \\ \dfrac{dR}{dt} = \gamma I, \end{cases} \tag{1}$$

where $N = S(t)+E(t)+I(t)+R(t)$, and the contact rate β, the latent rate σ, the recovery rate γ are model parameters to be determined.

The discrete format of equation (1) is as follows:

$$\begin{cases} S_{t+\triangle t} = S_t - \dfrac{\beta S_t I_t}{N} \cdot \triangle t, \\ E_{t+\triangle t} = E_t + (\dfrac{\beta S_t I_t}{N} - \sigma E) \cdot \triangle t, \\ I_{t+\triangle t} = I_t + (\sigma E_t - \gamma I_t) \cdot \triangle t, \\ R_{t+\triangle t} = R_t + \gamma I_t \cdot \triangle t. \end{cases} \tag{2}$$

3 Infection Prediction via LSTM

In this section, we aim to predict the cases of infection with LSTM [14]. The input X of LSTM is the data of 5 consecutive days, $X = \{x(t), x(t+1), ..., x(t+4)\}$, where $x(t) = \{S(t), E(t), I(t), R(t)\}$ and $S(t)$, $E(t)$, $I(t)$, $R(t)$ denotes the number of susceptible, exposed, infected and recovered individuals of day t respectively. The output is the prediction value of $\{S, E, I, R\}$ for the following day 6.

The structure of our model is shown in Fig. 1. The input data X goes through LSTM layer with 16 hidden units, and the output of the last LSTM cell passes through a fully connected layer to yield the four-dimensional prediction value $\{S, E, I, R\}$ for the next day.

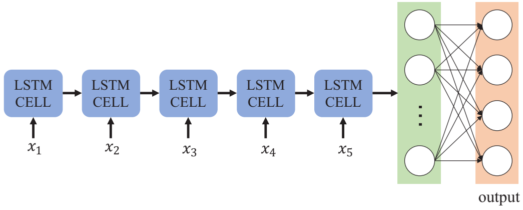

Fig. 1: The structure of LSTM model.

Before training the model, we apply min-max normalization to the data to constrain the values of the data between 0 and 1. The data of last 10 and 20 days are split as the test set in prediction, while the rest is for the training set. We choose Mean Square Error (MSE), $MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y_i} - y_i)^2$ where $\hat{y_i}$ represents the estimation of a data point y_i, as the loss function. We use Adam optimizer and set the learning rate as 0.001. The LSTM model is trained for 1000 epochs with a batch size of 16.

For evaluation, the model is used to predict infection on the test set (days after the training set). The prediction is performed iteratively, we use last 5 days of data in the training set to predict the first day of the test set, then use last 4 days of data in the training set and the prediction of the first day on the test set to predict the second day of the test set and so on. We introduce the Mean Absolute Percentage Error (MAPE), $MAPE = \frac{100\%}{n}\sum_{i=1}^{n}|\frac{\hat{y_i}-y_i}{y_i}|$ with $\hat{y_i}$ represent the estimation of y_i, to quantify prediction accuracy. The MAPE values for the prediction of 10 and 20 days are 10.8% and 30.0% on average, respectively. The curves of the prediction and test set are shown in Fig. 2. Obviously, it can be observed that the LSTM technique can predict the short-term trend with a high accuracy, while failing to predict the long-term evolution.

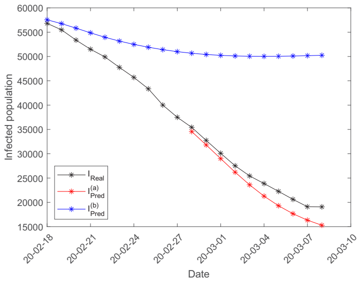

Fig. 2: The prediction of infected population via LSTM. (a) Prediction of 10 days. (b) Prediction of 20 days.

4 Parameter Identification of The Single-region SEIR Model

Although LSTM can make short-term predictions of epidemiological trends with a relatively high accuracy, it cannot explain the dynamics of the epidemic transmission process and is not suitable for predicting the long-term evolution of infected cases. Therefore, we further seek to identify the parameters in the SEIR model to test whether it is suitable to describe the spread trend. Suppose we have four-dimensional finite time series data $X = \{S(t), E(t), I(t), R(t)\}$ generated by the SEIR model. We can rewrite equation (2) into $X_{t+\triangle t} = \Phi(X_t)$, it means that $X_{t+\triangle t}$ can be obtained by X_t , if $\Phi(\cdot)$ is known. So our optimization objective function is:

$$\max_{\theta_{i-1}}\{P(\hat{X}(t+\triangle t) = X(t+\triangle t)|X(t))\}. \tag{3}$$

Therein, $\hat{X}$ represents the estimation of X. Based on this model and the data published earlier, we can use $\Phi(\cdot)$ to fit the data. However, in the SEIR model, the number of exposed individuals cannot simply be equal to the number of reported suspected cases, so the problem becomes identifying four parameters $\Theta = \{E(0), \beta, \sigma, \gamma\}$, and the optimization objective function is transformed into:

$$\min_{\Theta}\max_{\forall i}\left\{\frac{I(t_i,\Theta) - \hat{I}(t_i,\Theta)}{I(t_i,\Theta)}, \frac{R(t_i,\Theta) - \hat{R}(t_i,\Theta)}{R(t_i,\Theta)}\right\}. \tag{4}$$

Here, I and R represent published data of confirmed and recovered cases including deaths, respectively. Since the number of people in the previous period is small, so we take the maximum relative error of I and R as the objective function, and minimize it to obtain the parameter value.

To identify the unknown parameters and capture the initial value of the exposed number $E(0)$. We propose an online parameter identification method based on the idea of sliding window, and make the method forgetful to the past information by using the dynamic window size. At the same time, the forgetting effect can be adjusted by setting the last window size.

First, we calculate the predicted values in the sliding window with the given initial values $\hat{S}_1, \hat{E}_1, I_1, R_1, \theta_0$, and use the trust region reflective algorithm to minimize the maximum relative error as the goal to obtain the suboptimal parameter θ_1. Then slide the window to the second variable $\hat{S}_2, \hat{E}_2, I_2, R_2$ and use these values as the initial value with parameter θ_1 to predict the variables in the slide window, repeat the optimization method of the first step, So we can get the predicted $\hat{S}_i, \hat{E}_i$ values and the parameter θ_{i-1} that optimize the prediction results, Repeat the above steps until there are no more known I_i, R_i on the right of the side window, Here, we can use the predicted value $\hat{S}_i, \hat{E}_i$ and the true data I_i, R_i in the window with parameter θ_{i-1} to predict the variables on the right of the slide window. However, doing so ignores the variable information in the window, Therefore, in order to make full use of limited time information, we choose to make the window smaller and make the next prediction based on the information in the new window to update the variables and parameters. The advantage of this method is that the prediction of the future can be based on the newer information and the prediction function has a robust prediction ability.

In each slide window the objective function is defined as:

$$\max_{\Theta}\{P(\hat{X}_j = X_j)|X_i(t)), X_j \in Window\} \tag{5}$$

The illustration of our method is shown in Fig. 3.

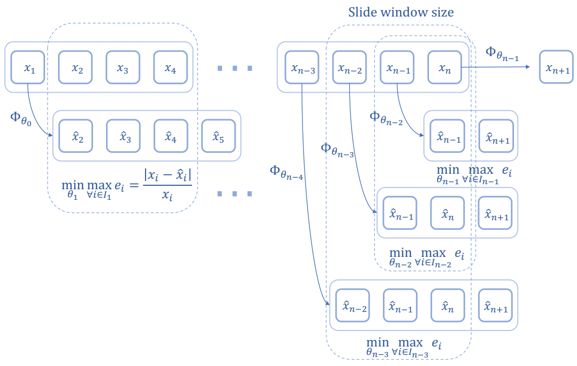

Fig. 3: The structure of the sliding window method. Therein, I_i denotes the ith window.

When the parameters change slowly, the influence of the previous data on the model as a whole is gradually reduced after the forgetting effect is added. At the same time, for the addition of new data, the method can dynamically adjust the model parameters based on the past information to modify the future prediction direction of the model, which not only enhances the prediction ability, but also reduces the calculation complexity.

5 Numerical Experiment

Based on the proposed method above, we conducted numerical experiments with the COVID-19 data in China. We set the tunable parameter sliding window size as 7, the last window size as 3 and initialize the parameters to be identified between 0 and 1. Note that, generally, with a larger window size, the sensitivity of the prediction becomes weak. However, the smaller window size will result in an overfitting prediction.

The training data set of the country starts from January 10th to February 16th, and the test set data starts from February 17th to March 8th, including numbers of the current confirmed cases, recovered patients and death. The estimated exposed cases $\hat{E}_1$ is initialized between 8 to 56 when the step size is 1, and the population N is 1,400,000,000. The optimization is accomplished with the trust region reflective algorithm [15].

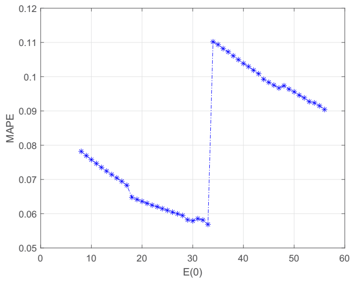

Fig. 4: MAPE with different values of E(0) of the test set.

For evaluation, same as LSTM, the model is used to predict the infection cases of the test set. We use the data and parameter θ in the last day of the training set to predict the situation of following 21 days. The MAPE of the 21-day prediction is 8.15% on average, see Fig. 4. The parameter identification results of different E(0) and days on training set are shown in Fig. 5.

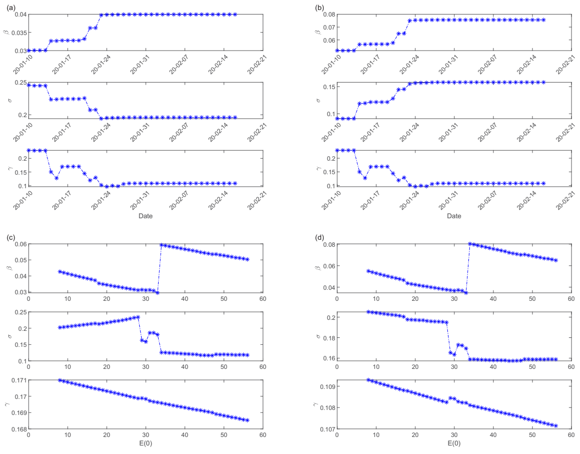

Fig. 5: Parameter identification with different values of $E(0)$ on different days. (a) $E(0) = 24$. (b) $E(0) = 40$. (c) Parameters for January 20th. (d) Parameters for February 1st.

Then, we set the sliding window size and the last window size as 2, to make the most use of the data. We analyze the situation of China and make prediction by our sliding window model. For comparison, the curve of the prediction and test set are shown in Fig. 6. The MAPE for the prediction of 21 days is 3.88%. The forecast result suggests that we can predict the inflection point of the curve and COVID-19 will cease in the end of April in China.

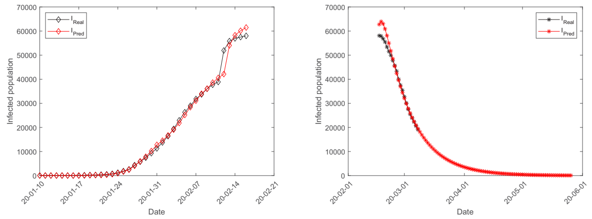

Fig. 6: Prediction of the infected population. Left figure shows the prediction on training set, while the right figure indicates the prediction performance.

6 Conclusion

In order to study the transmission dynamics of covid-19 and help people understand its evolution trend, this paper first use an LSTM model to predict the infected population in China. The MAPE results have shown that it is accurate for short-term prediction, but not for long-term evolution. In addition, because it does not explain the dynamic strategy of the spread process, we introduce the widely accepted SEIR model to describe the epidemic spread process. The deterministic SEIR model may capture the dynamics of the epidemic spreading process. By using a sliding window method, with a proper selection of the parameters, we suggest that we may predict the accurately the trend of the populations in different cases. By using a small window size in the method, we may sensitively acquire the inflection of the trend of COVID-19 and predict the endpoint of COVID-19 in different provinces and countries.

In short, the actual spreading process of COVID-19 shall be much more complicated, which cannot be entirely captured by the limited data and any theoretical model. In future work, we will extend the current SEIR model by considering the mobility of population for different sub-regions in China from a network perspective, and further investigate the general dynamics of COVID-19 and the general spread strategies of more epidemic viruses.

References

[1] S. Funk, E. Gilad, C. Watkins and V. A. Jansen, The spread of awareness and its impact on epidemic outbreaks, *Proceedings of the National Academy of Sciences*, 106(16): 6872-6877, 2009.

[2] D. S. Merrell, S. M. Butler, F. Qadri, N. A. Dolganov, A. Alam, M. B. Cohen, S. B. Calderwood, G. K. Schoolnik and A. Camilli, Host-induced epidemic spread of the cholera bacterium, *Nature*, 417(6889): 642-645, 2002.

[3] M. J. Keeling and K. T. Eames, Networks and epidemic models, *Journal of the Royal Society Interface*, 2(4): 295-307, 2005.

[4] C. N. Angstmann, A. M. Erickson, B. I. Henry, A. V. McGann, J. M. Murray and J. A. Nichols, Fractional order compartment models, *SIAM Journal on Applied Mathematics*, 77(2): 430-446, 2017.

[5] D. Brockmann and D. Helbing, The hidden geometry of complex, network-driven contagion phenomena, *Science*, 342(6164): 1337-1342, 2013.

[6] L. Hufnagel, D. Brockmann and T. Geisel, Forecast and control of epidemics in a globalized world, *Proceedings of the National Academy of Sciences*, 101(42): 15124-15129, 2004.

[7] P. Bajardi, C. Poletto, J. J. Ramasco, M. Tizzoni, V. Colizza and A. Vespignani, Human mobility networks, travel restrictions, and the global spread of 2009 H1N1 pandemic, *PLoS One*, 6(1), 2011.

[8] X. Wang, W. Ni, K. Zheng, R. P. Liu and X. Niu, Virus propagation modeling and convergence analysis in large-scale networks, *IEEE Transactions on Information Forensics and Security*, 11(10): 2241-2254, 2016.

[9] C. Fraser, C. A. Donnelly, S. Cauchemez, W. P. Hanage, M. D. Van Kerkhove, T. D. Hollingsworth, J. Griffin, R. F. Baggaley, H. E. Jenkins and E. J. Lyons, Pandemic potential of a strain of influenza A (H1N1): early findings, *Science*, 324(5934): 1557-1561, 2009.

[10] J. Wang, X. Wang and J. Wu, Inferring metapopulation propagation network for intra-city epidemic control and prevention. in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018: 830-838.

[11] W. Wang, Q. H. Liu, J. Liang, Y. Hu and T. Zhou, Coevolution spreading in complex networks, *Physics Reports*, 2019.

[12] A. Koher, H. H. Lentz, J. P. Gleeson and P. Hövel, Contact-based model for epidemic spreading on temporal networks, *Physical Review X*, 9(3): 031017, 2019.

[13] M. Y. Li and J. S. Muldowney, Global stability for the SEIR model in epidemiology, *Mathematical Biosciences*, 125(2): 155-164, 1995.

[14] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural computation*, 9(8): 1735-1780, 1997.

[15] Y. Yuan, Recent advances in trust region algorithms, *Mathematical Programming*, 151(1): 249-281, 2015.