2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE) | 978-1-7281-9146-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICAICE51518.2020.00021

Machine Learning Methods on COVID-19 Situation Prediction

Zhihao Yang[1, a, †] and Kang'an Chen[2, b, †]
[1]School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, 100876, China;
[2]School of Mathematic and Statistics, Wuhan University, Wuhan, Hubei 430072, China.
[a]waldoyzh@bupt.edu.cn, [b] 2017301500068@whu.edu.cn
[†]These authors contributed equally.

***Abstract*-COVID-19 has already had a devastating impact on economic and social development and people's life all over the world. Up to September 22nd, 2020, more than 30 million people have been infected. Finding out how to predict and estimate the pandemic trend precisely is of huge necessity because COVID-19 has made the world economy in a recession and deprived over 700 thousand lives. This paper is dedicated to making a comparison between conventional machine learning regression models, including ridge regression and lasso regression, and multivariate polynomial regression. Besides, we attempt to use statewide data to fit nationwide data in the US, proposing a novel aspect to forecast pandemic trend.**

Under the current situation, insufficient data limit the use of machine learning. To address the issue of data deficiency, a classification model based on neural network with Twitter data is applied. This gives out an alternative approach to estimate daily increase of infected people.

We discovered that in a different period, specific models would outweigh another models' performance. Also, our result showed that the data of Georgia and Massachusetts could represent the whole nation data with linear transformations. And this paper verified that using alternative data that relate to the COVID-19 situation to alleviate data deficiency is feasible.

Keywords: COVID-19, Machine learning, Trend prediction, Regression model, Neural network

I. INTRODUCTION

Since January 2020, the COVID-19 pandemic has put the world into a tough situation. The world economy, people's daily life is severely affected, including a slump in the stock market, interruption on education activities [1,2,3]. Even for now, the death toll and infected people are still soaring up in most countries over the world.

It is a dilemma for our society to balance economic growth and pandemic prevention. Hence, in order to control the spread of COVID-19, this unprecedented public health crisis, it is of a great necessity to predict the trend of the epidemic as precise as possible, providing valuable references to policymaking, and resource allocating.

Predictions based on machine learning models have obtained their popularity in recent years. For the COVID-19 prediction, Ribeiro, M. H. D. M., et al., implemented support vector regression and stacking-ensemble, etc. with clinical data and reached an error in the range of 0.87%-3.51% with one day ahead [4]. Chimmula, V.K.R., et al. developed a forecasting model using long-short term memory (LSTM) networks with data from John Hopkins University to estimate the ending point of the COVID-19 in Canada [5]. The model estimated that this outbreak would terminate around June 2020, in Canada. By unveiling the significant features, their work was plausible in predicting the pandemic trend. Peipei Wang., et al. utilized the Logistic model to fit the paramount points of the outbreak and fed the values into the FbProphet model, a time series model, to predict the pandemic trend. It showed that the global peak would occur in late October, with 14.12 million people infected cumulatively [6].

However, since the outbreak of pandemic beginning in March globally, and this led to the deficiency in direct related data onto COVID-19, qualifying the reliability of forecasting results [7]. And most of the existing works on forecasting were from a nationwide perspective, which employed nationwide data to predict a nationwide trend. This may occlude other approaches to address the forecasting problem. To address this problem, a primary issue is whether indirect data is capable of containing valuable information about pandemic. The pandemic has affected people in many ways, not only on the clinical aspect but also on people's social behaviors, like the attitude behind their tweets. If these features are applied to the prediction model, the model might be more robust.

Considering the drawbacks, the paper is dedicated to contrasting conventional machine learning regression models to predict pandemic trends in the US, providing references to other scholars to do further researches. Meanwhile, we believe introducing new perspectives of data may bring better model performance for us, enlarging dataset size to obtain a more precise distribution. Therefore, we attempt to use statewide data to fit nationwide data in the US, proposing a novel aspect to forecast pandemic trend. Furthermore, we try to build a classified neural network based on tweets with hashtags about pandemic to estimate daily infected people in the US, giving out an alternative approach to avoid data deficiency. As far as we know, only a few previous researches have investigated indirect data like social media to fit the trend.

It is noticed that on this specific issue, different models vary on performance. For each stage, some models fit better than others. Also, we find that data for Georgia and Massachusetts can faithfully represent the US nationwide trend with linear transformation, and emotional analysis for tweets reveals a negative related trend on pandemic trend. Once this trend is

DOI 10.1109/ICAICE51518.2020.00021

revealed, employing it to predict the pandemic situation in the following days and contribute to policymaking is possible.

The rest of this paper is structured as follows. In Section 2, the paper mainly discusses our methods. Experiment results and analysis are in Section 3. Section 4 contains the conclusion of our work. And in Section 5, we briefly discuss the future challenges.

II. METHOD

In this section, we will elaborate on the models used in our paper, which can be divided into two kinds. One is traditional regression models, including Lasso Regression, Ridge Regression, and Multivariate Polynomial Regression. Another is the models based on neural networks, including FCNN (Fully-Connected Neural Network), LSTM (Long Short-Term Memory), and a classification model built by us. This classification model bases on twitter sentimental analysis and uses the overall percentage of different emotion types to predict a plausible range of daily increase numbers.

A. Regression Models

Traditional regression models fade in recent years due to the popularity of the neural network. However, for consecutive value-based predictions, regression models are still robust [8].

Lasso Regression, Ridge Regression, and Multivariate Polynomial Regression are the three models used in the paper. All the models are based on the linear regression model.

The linear model is a mainstay of statistics. Given a vector of inputs $X^T = (X_1, X_2, \cdots, X_p)$, one can estimate the output Y via the model

$$\hat{Y} = \hat{\beta}_0 + \sum_j^p X_j \hat{\beta}_j \tag{1}$$

The term $\hat{\beta}_0$ is the intercept. It is convenient to include the constant variable 1 in X , include $\hat{\beta}_0$ in the vector coefficients $\hat{\beta}$, and then write the linear model in vector form as an inner product

$$\hat{Y} = X^T \hat{\beta} \tag{2}$$

where X^T denotes vector or matrix transpose. There are many different methods to fit the linear model to a set of training data, and the basic one is the least squares. In this approach, it is coefficients β picked to minimize the residual sum of squares

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 \tag{3}$$

$\mathrm{RSS}(\beta)$ is a quadratic function of the parameters, and hence its minimum always exists, but may not be unique. The solution is easier to characterize in matrix notation. It can be written as Equation 4

$$\mathrm{RSS}(\beta) = (y - X\beta)^T (y - X\beta) \tag{4}$$

where X is a $N \times p$ matrix with each row an input vector, and y is an N-vector of the outputs in the training set. After differentiating w.r.t. β , a normal equation is conducted as Equation 5

$$X^T (y - X\beta) = 0 \tag{5}$$

If $X^T X$ is nonsingular, then the unique solution is given by

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{6}$$

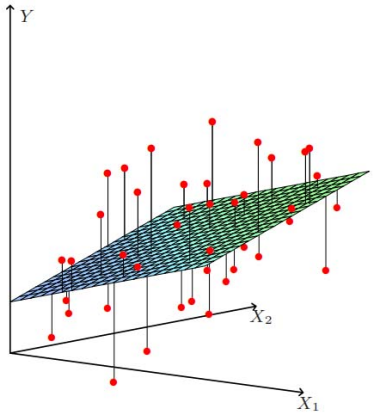

Figure 1. An example of a 2-dimension linear regression model [9].

1) Ridge Regression

Ridge regression shrinks the regression coefficients by imposing a penalty on their size. The ridge coefficients minimize a penalized residual sum of squares. It is almost the same with the linear model except for changing the criterion to choose $\hat{\beta}$:

$$\mathrm{RSS}(\lambda) = (y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta \tag{7}$$

Here $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage: the larger the value of λ is, the greater the amount of shrinkage can be.

By minimizing $\mathrm{RSS}(\lambda)$, the solution is

$$\hat{\beta}_{\mathrm{ridge}} = (X^T X + \lambda I) X^T y \tag{8}$$

2) Lasso Regression

The lasso is another shrinkage method like a ridge, with subtle but important differences. The lasso estimation is defined by

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \} \tag{9}$$

3) Multivariate Polynomial Regression

The third model, the multi-polynomial regression model, shares the same basic model and criterion to choose $\hat{\beta}$ with linear regression. The only difference is that this model creates some new features by calculating the powers or the cross product of the data.

B. Neural Network Models

The neural network is the most important role in machine learning nowadays. By using a neural network, it is possible to extract abstract features from original data, and therefore, meet our goal to predict the pandemic situation in a new way.

A fully connected neural network is the most basic and intuitive one. The basic component of it is the neuron. Neurons can accept inputs from previous neurons in the previous layer and perform non-linear transformations with an activation function. With all neurons connected with proper structure, weights, and bias, a sophisticated non-linear function is generated. Theoretically, a neural network is capable of fitting any complex function with infinite neurons; therefore, it can perform any task. However, better performance and less time to calculate are irreconcilable. Designing networks with a specific feature to balance performance and time consumption is necessary.

LSTM, a "state-of-the-art" model to predict time series data. The most important feature of it is that it combines previous output with current input to feed the neurons. So, LSTM includes a time window in it. This underscores either gradient vanishing or gradient explosion in long time series training [10]. A typical LSTM neuron is shown in Figure 2.

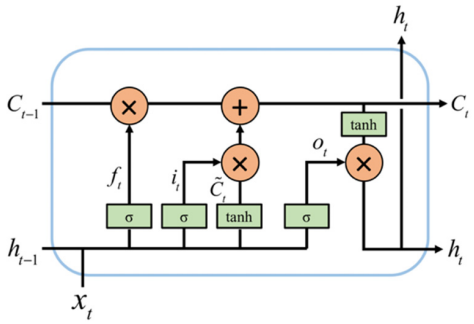

Figure 2. A typical LSTM neuron

LSTM is capable of learning long term dependencies. This feature depends on a special structure called “Cell state” or a rather sophisticated gate control system. Equation 10 to Equation 12 are input gate, forget gate, and output gate, respectively.

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{10}$$

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{11}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{12}$$

The forget gate is a layer with sigmoid to control whether previous information will be feed in the current neuron (remember or abandon).

The input gate decides current information. Sigmoid selects which information would be updated, and tanh creates new candidate memory to be selected by sigmoid.

The output gate gives out value to feed the following neurons.

With these gates, LSTM can generate “long memory” and “short memory” with Equation 13 and Equation 14 for a long or short memory, respectively.

$$C_t = f_t C_{t-1} + i_t \tilde{C}_t \tag{13}$$

$$h_t = o_t \tanh(C_t) \tag{14}$$

This offers the network capability to solve the long-term dependency, maximizing the value of limited data in a long range.

Tweets analysis is based on Textblob (https://github.com/sloria/TextBlob) and neural network. Textblob is used to analyze sentimental information behind the natural language.

The neural network is based on fully connected dense layers. Furthermore, we implement dropout layers inside the network, between dense layers. Dropout layers help to avoid overfitting during training; therefore, it increases the generalization ability of models [11]. The network structure is shown in Figure 3.

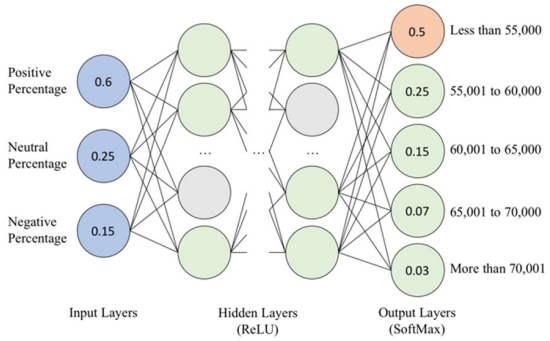

Figure 3. Classification network model

The essential idea of tweets analysis is using the relationship between latent emotional attitude and increase the infected number, which is how people respond to a different number. For instance, people intuitively feel more positive in the infected number was 1,000 other than 10,000. Hence, the percentage of positive attitude, negative attitude, and neutral attitude may somehow represent the pandemic situation.

III. Analysis and Discussion on Results

A. Data Source and Preprocessing

Three separate portions of our data are applied in experiments, including US nationwide pandemic data, US statewide pandemic data, and Twitter data.

These data are employed to train, validating and testing the regression model. Also, our model, which fits the state data to national data, depends on these data. The important properties of the dataset are listed in Table 1.

TABLE I. DATA INFORMATION

Type	Size	Features	Source
US Nationwide	Mar 15th – Aug 7th, daily	daily infected, death toll, on ventilator increase	John Hopkins Coronavirus Resource Center and Worldometer
US Statewide	Mar 15th – Aug 7th, daily	daily infected, death toll, on ventilator increase	John Hopkins Coronavirus Resource Center and Worldometer
Tweets sent from the US	28,263 Tweets	With hashtags of #covid, #covid-19, #coronavirus, #pandemic, #epidemic	Twitter

In the original data, early stage data are not available. This might cause by the ignorance of pandemic spread. We inspect missing data, noticing that these data are close to 0 from the start. Hence, we fill the blanks with 0 or 1 to make the values consecutive and plausible.

After inspecting nationwide and statewide data, we notice that an increased infected number shows three apparent time spans. Hence, using increase features to build models may outvie the cumulative features.

Then, we split the data into three different temporal portions by local maximum or local minimum: from Mar 15th to Apr 15th, from Apr 16th to Jul 15th, and from Jul 16th to Aug 7th. This will help us to deploy different models on each span to investigate model performance. Figure 4 represents the whole trend from Mar 15th to Aug 7th.

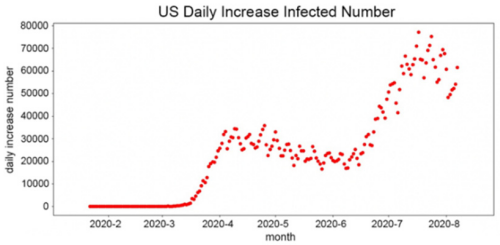

Figure 4. US daily increase number

For Twitter data, we collect tweets with hashtags aforementioned by date. Considering the relativity of tweets and pandemic situations, finding a quantitative relationship is not feasible. Therefore, we use Textblob to run sentimental analysis on these tweets and categorize tweets into three categories: positive, negative, and neutral. Then calculate the ratio of each kind, preparing for the classifier training process.

B. Experiment Result

In Table 2, we display MSE results of lasso regression, ridge regression, and multivariate polynomial regression on all three-time spans.

TABLE II. MSE RESULT ON EACH SPAN WITH MODELS

	Span 1	Span 2	Span 3
Lasso	1,796,181.54	7,368,820.97	56,214,515.58
Ridge	1,903,555.59	7,368,820.97	56,191,697.69
Mul-Poly	7,264,723.99	8,799,535.36	18,642,735.30

In the first span, lasso regression wins out with the lowest MSE; lasso regression and ridge regression tie but still beat multivariate polynomial regression in span 2. However, in the last span, multivariate polynomial regression gets the minimum MSE. It is obvious that on each span, not all models work well. Using lasso, ridge (or lasso), and multivariate polynomial regression on each span respectively to predict the pandemic data is more plausible.

Figure 5 and Figure 6 represent our work on statewide data fitting. Figure 5 contains the daily increase number of infected people in Georgia and Massachusetts with a red line and green line, respectively. Figure 6 is a synthesized line with weighted data from Georgia and Massachusetts combined linearly.

The weights are 13.67 for Georgia and 2.86 for Massachusetts. It is of a high similarity in comparison with real national data scatter in Figure 3.

It is noticed that the daily increase infected in Massachusetts reached a local maximum in April and Georgia did so in July, which correlates two local maximums in US national data precisely. Hence, employing state data to estimate the whole nation's data is feasible. (A surge in Massachusetts around June 1st was caused by the alternation of accounting standards by Massachusetts State Government. All confirmed infected from March 1st were contained in data of June 1st).

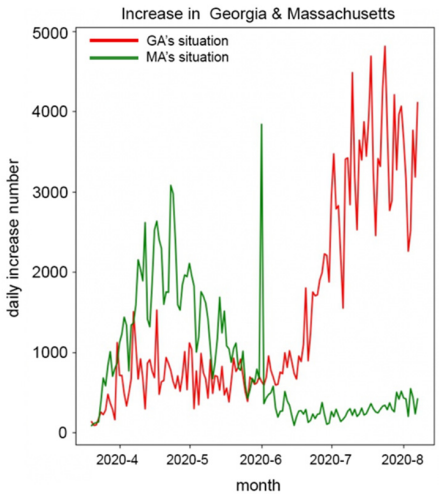

Figure 5. Daily increase in two states

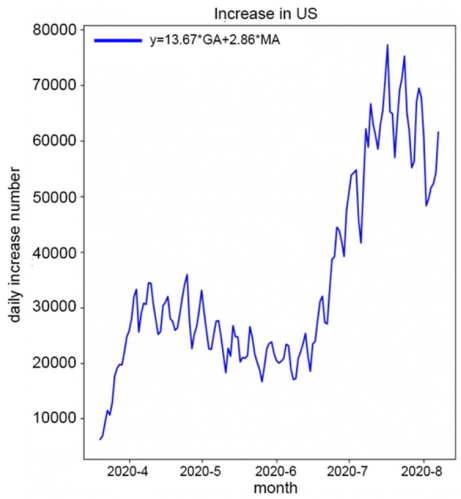

Figure 6. Estimation of daily increase in the US

With this, it is possible to predict a nationwide pandemic situation way before the whole nation data is collected and analyzed, boosting the policy making process and save more lives.

Since results from FCNN and LSTM do not meet our expectations due to data deficiency, with the highest accuracy of 37%, exploring new approaches to make the network more robust is significant. Rather than optimizing network structure, introducing more data to train network is more feasible.

In Twitter sentimental analysis-based classification, we define five ranges for daily increase infected number, ranging from 55,000 to 70,000. Range 1 contains all situations below 55,000; range 2 ranges from 55,001 to 60,000; range 3 starts at 60,001 and ends at 65,000; range 4 starts at 65,001 and ends at 70,000, and range 5 is above 70,001. Our training on this network converges at around 50 epochs with a validating accuracy of 67%. Then we test a set of real data with an input of (.329, .222, .449). Numbers in parenthesis denote positive ratio, negative ratio, and neutral ratio, respectively. The output probability vector is (.0866, .1577, .2779, .3163, .1614), which indicates this situation should be classified into range 4. The actual number of that is 66,969, which also lies in range 4. With a higher ratio of negative attitudes, the number of infected people decreases. This verifies that some non-medical data are having a relationship with the pandemic situation and might be used to feed the networks to make predictions.

IV. Conclusion

In summary, our paper simulated the epidemic trend onto the US by using conventional machine learning regression models, including lasso regression, ridge regression, and multi-polynomial regression. In the time spans divided above, we find that lasso regression owns the lowest MSE in the first span; in the second span, MSE of lasso regression and ridge regression tied for the minimum; in the last span, multivariate polynomial regression reaches the minimum MSE. We use a lasso, ridge (or lasso), and multivariate polynomial regression on each span, respectively, to stimulate the pandemic data.

Meanwhile, we fit nationwide data in the US by using statewide data. We notice that Massachusetts reached a local maximum in April, and Georgia did so in July, which correlates two local maximums in US national data precisely, in which case can we forecast the pandemic trend of the whole country through the data of only two states.

Besides, we also use neural network models, including FCNN, LSTM, and neural network based on twitter sentimental analysis to simulate the epidemic trend in the US. FCNN and LSTM, with the highest accuracy of 37%, do not have a satisfying result due to data deficiency. In Twitter sentimental analysis-based classification, we define five ranges for daily increase infected number and predict daily increase infected number on a chosen day using the positive ratio, negative ratio, and neutral ratio in three categories we divided above. The analysis result shows that attitude ratio is negatively related to infected people's numbers. Furthermore, this paper indicates that it is possible to forecast the pandemic trend using data collected from other perspectives.

V. Discussion

This paper focuses on the simulation and prediction of the COVID-19 daily positive increasing numbers. The basic idea is to excavate the relationship between data of several days in a row. However, a daily positive increasing number is influenced by many factors. For instance, when it comes to a discussion about herd immunity, the efficiency of government implementation and environmental factors, like people's attitude towards COVID-19 or people's tolerance of death toll, will be the crucial elements to influence or restrict the increase of positive increasing number. One may build a better model through adding new features, which can reflect government implementation, local culture, and environmental factor. However, this kind of new features may not be quantified easily.

Another way to deal with these factors which cannot be added to the existing model simply is to use constraints, and we may even create a model merely using constraints. For example, SEIR is a kind of epidemic model, which is a set of differential equations and has four hyperparameters. We attempt to apply this model on the data and solve the partial differential equations through a package in Python, named neurodiffeq. However, this package cannot solve the nonlinear equation without major modifications. This may attract more researchers to contribute valuable researches in the future.

The proposal of Twitter analysis is an attempt to settle the problem of small sample learning. There are two methods to solve the problem of a small data size. One is boosting, in other words, generating new data from the present data. This method can improve the accuracy; however, the promotion will be limited because the generated data will form the same distribution as the original data. Another is meta-learning, a distinctive way to classify or predict the giving test set through other datasets. For data augmentation, a more precise distribution is much more important than larger data size. How to fit the actual distribution while generating data based on current distribution is still to be discussed.

REFERENCE

[1] Atkeson A. What will be the economic impact of covid-19 in the us? rough estimates of disease scenarios. National Bureau of Economic Research; 2020 Mar 19.

[2] Sintema, Edgar John. "Effect of COVID-19 on the performance of grade 12 students: Implications for STEM education." Eurasia Journal of Mathematics, Science and Technology Education 16.7 (2020): em1851.

[3] van Dorn, Aaron, Rebecca E. Cooney, and Miriam L. Sabin. "COVID-19 exacerbating inequalities in the US." Lancet (London, England) 395.10232 (2020): 1243.

[4] Ribeiro MH, da Silva RG, Mariani VC, dos Santos Coelho L. Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. Chaos, Solitons & Fractals. 2020 May 1:109853.

[5] Chimmula VK, Zhang L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos, Solitons & Fractals. 2020 May 8:109864.

[6] Wang P, Zheng X, Li J, Zhu B. Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. Chaos, Solitons & Fractals. 2020 Oct 1;139:110058.

[7] Ardabili SF, Mosavi A, Ghamisi P, Ferdinand F, Varkonyi-Koczy AR, Reuter U, Rabczuk T, Atkinson PM. Covid-19 outbreak prediction with machine learning. Available at SSRN 3580188. 2020 Apr 19.

[8] Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. Introduction to linear regression analysis. Vol. 821. John Wiley & Sons, 2012.

[9] Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. New York: Springer series in statistics, 2001.

[10] Gers, Felix A., Jürgen Schmidhuber, and Fred Cummins. "Learning to forget: Continual prediction with LSTM." (1999): 850-855.

[11] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research. 2014 Jan 1;15(1):1929-58.