2020 International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI) | 978-1-6654-2316-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICHCI51889.2020.00080

U.S. Pandemic Prediction Using Regression and Neural Network Models

Tianxiao Liu
Dalian Yuming Senior High School
Dalian, China
worldfly1@hdu.edu.cn

***Abstract*—With the global outbreak of COVID-19 in 2020, it is essential for government to make aware of the trend of the pandemic. To achieve this goal, some regression and neural network models are used to predict pandemic data of the U.S. Three models -linear regression, logistic regression, and Recurrent Neural Network (RNN)- are selected for predicting cases per million people in America. Then, the effectiveness of these models is compared. These models are evaluated using Mean Squared Error (MSE). It can be concluded that while the traditional regression models, including linear and logistic regression, are much more efficient for inference, RNN predicts more accurately, with the smallest MSE being nearly 2.8. This paper gives effective guidance for American governments on how to select models to predict relevant data of the pandemic.**

Keywords- Pandemic Prediction; Linear Regression; Logistic Regression; RNN; COVID-19

I. INTRODUCTION

Due to the global pandemic of COVID-19 in 2020, an effective model for predicting the trend of this virus is of vital necessity. Some useful methods can be learned from the previous studies. To this end, it is essential for researchers to find an effective model for making prediction of COVID-19. We can learn from previous experience in SARS, a kind of similar pandemic in China in 2003. Elaine et al. [1] aimed to use a combination of simulations and classification techniques to predict epidemic curves and infer underlying disease parameters for an ongoing outbreak. Six supervised classification methods were used in identifying partial epidemic curves from six agent-based stochastic simulations of influenza epidemics. Wang et al. [2] integrated the most updated COVID-19 epidemiological data before June 16, 2020 into the logistic model to fit the cap of this epidemic trend, and then fed the cap value into the FbProphet model, a machine learning based time series prediction model to derive the epidemic curve and predict the trend of the epidemic. Three significant points were summarized from their modeling results for global, Brazil, Russia, India, Peru, and Indonesia. Under mathematical estimation, the global outbreak will peak in late October, with an estimated 14.12 million people infected cumulatively. Nan and Gao's research [3] used search engine data to monitor and forecast AIDS in China. A machine learning-based method - artificial neural networks (ANNs)- was used to forecast AIDS incidences and deaths. The AIDS trend data from the largest Chinese search engine, Baidu.com, were collected and selected as the input variables of ANNs, and the officially reported actual AIDS incidences and deaths were used as the output variable. Three criteria - the mean absolute percentage error, the root mean squared percentage error, and the index of agreement - were used to test the forecasting performance of the ANN. Based on the monthly time series data from January 2011 to June 2017, Nan and Gao found that, under these three criteria, the ANN method could lead to satisfactory forecasting of AIDS incidences and deaths, regardless of the change in the number of search queries. Despite the inability to self-detect HIV/AIDS through online searching, Internet-based data should be adopted as a timely, cost-effective complement to a traditional AIDS surveillance system. Their objectives were to investigate the utility of boosted regression trees (BRTs) [4] for predictive modeling of FHB epidemics in the United States, and to compare the predictive performances of the BRT models with those of logistic regression models that had been developed previously. Shah et al. [4] led to novel insights into the weather–epidemic relationship. Nsoesie et al. [5] introduced an approach for inference on infectious disease data based on supervised learning. They extended this work to the case where the underlying infectious disease model was inherently spatial. Their goal was to compare the use of global epidemic curves for building the classifier, using spatially stratified epidemic curves. The results of the simulated data proved the effectiveness of this method. One can refer to [6-10] and references therein for more details.

In this study, the COVID-19 pandemic in the U.S. is analyzed and predicted using regression and neural networks. First, since the data are with much noise and are incomplete, the raw data is preprocessed [11]. Then, three latest models are used, including linear regression, Logistic regression, and Recurrent Neural Network (RNN), to analyze and predict the pandemic, and the effectiveness of these models is compared. These values are evaluated with Mean Squared Error (MSE), and it can be concluded that the best model is RNN, with the smallest MSE. It is expected that work can help the government to manage the public resource and can help medical workers to supply the drugs.

II. METHOD

A. Data Description

Data preprocessing is first applied to the raw data, which is incomplete, inconsistent, and with noise [12]. There are four steps in the data preprocessing: data cleaning, data transformation, feature selection, and feature extraction. Data cleaning is used to deal with incomplete data, and the data are deleted or filled in manually [13]. In the next step, the data are transformed by sampling them between 0 and 140 days after

978-1-6654-2316-8/20/$31.00 ©2020 IEEE
DOI 10.1109/ICHCI51889.2020.00080

the beginning of the pandemic in America and converting all of the data into integer form. Finally, the most relevant attributes and features, i.e., daily growth and total cases in America, are selected, and then the relevant data are extracted. This is because during specific analysis, there may be many attributes, but some attributes are irrelevant, and some are repetitive.

B. Linear Regression

Linear regression is a statistical analysis method that uses regression analysis in mathematical statistics to determine the quantitative relationship between two or more variables [14]. The expression form is y = Wx+e, and e is the normal distribution with a mean value of 0.

For linear regression, the data are split into two sets: train and test. And then, the Polynomial Degree Function is utilized to fit the model, which determines the greatest exponent of the function of regression [15]. Finally, the smallest MSE is located.

C. Logistic Regression

Logistic regression is similar to linear regression. Their main difference is the data type of y. The dependent variable y of linear regression analysis belongs to quantitative data, while the dependent variable y of logistic regression belongs to classified data.

Similar to linear regression, the data are split into the training set and the test set, and then the logistic regression is applied to the data [16]. Finally, the model is tested via cross-validation, which is a model validation method in machine learning.

D. RNN

The idea behind RNN is to use sequential information [17]. In traditional neural networks, it is assumed that all inputs (including outputs) are independent of each other. For many tasks, this is not a very good assumption. In order to predict the next word in a sequence, it is necessary to know which words precede it. RNN is circular because it performs the same operations for each element in the series, each of which depends on previous calculations. Another way is that RNN remembers the information that has been calculated so far. Fig. 1 shows the structure of RNN after being expanded into a whole network. The meaning of expanding here is expressing the network structure for the whole sequence. For example, if being concerned with a sentence with five words, the network will be expanded into a five-layer network with one later for each word.

In order to use RNN to predict the pandemic, it is necessary to use training RNN. Similar to the training process of a traditional neural network, a back-propagation algorithm is also used [18]. However, there will be some changes. Because the parameters are shared at all times of the network, the gradient output of each time depends not only on the calculation results at the current time, but also on the calculation results at all previous times. For example, in order to calculate the gradient at time t = 4, it is necessary to use back-propagation in 3 steps and add all previous gradients [19]. This is called Back Propagation over Time (BPTT) [20].

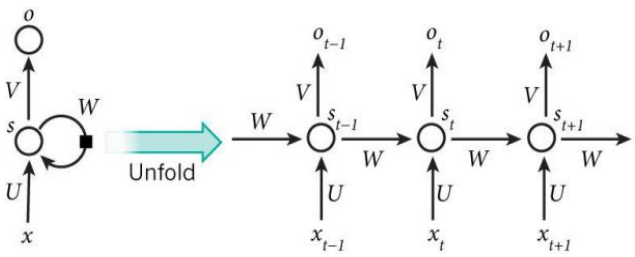

Figure 1. The illustration of RNN, which is widely used for sequence data prediction.

III. Results

This section discusses the advantages of the selected model. As shown in Table 1, their performance with aspect to accuracy and speed is summarized and the best model in accuracy and speed is concluded.

TABLE I. The overall comparison between the selected models according to accuracy and speed

Method	Linear regression	Logistic regression	RNN
MSE	9.0	100.0	2.8
Inference time	0.37s	0.43s	3.89s

A. Results for Linear Regression

With linear regression, the pandemic in America can be quickly and easily predicted with only several steps. The performance mainly depends on the complexity of data, where MSE is so large, especially when the relationship is not very linear, and there are big differences between the actual value and the predicted value. Fig. 2 shows that in some of the days between 100 and 140 after the start of the pandemic in America, the actual value of cases per million people is much larger than the predictive value of that. In Fig. 2, the MSE is just about 9.0, but if the original data becomes nonlinear in the future, the prediction will be even less accurate, and the MSE will be larger. It can be found that the linear regression model is the most efficient model among the competitors whose inference time is 0.37s. The relationship between observation and actual data using the linear model is shown in Fig. 2.

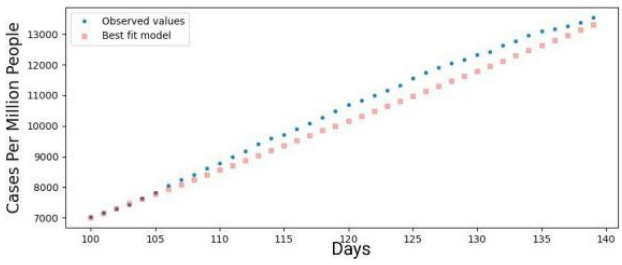

Figure 2. Regression results for linear regression

B. Results for Logistic Regression

With the same step as linear regression, the process using logistic regression is simple. Compared with linear regression,

Logistic regression gets a worse performance, with a larger MSE (MSE = 100.0). For example, Fig. 3 shows cases per million people between 100 and 140 days after the start of the pandemic in America. Since the best fit model of logistic regression has only one value between 100 and 140 days, the MSE is really large, and it cannot predict well in a later time. In Fig. 3, the MSE is about 100. It can be easily imagined that MSE can become even much larger in the future. The inference time of logistic regression is 0.43s, which is also a regression model with a fast speed. The relationship between observation and actual data using Logistic model is shown in Fig. 3.

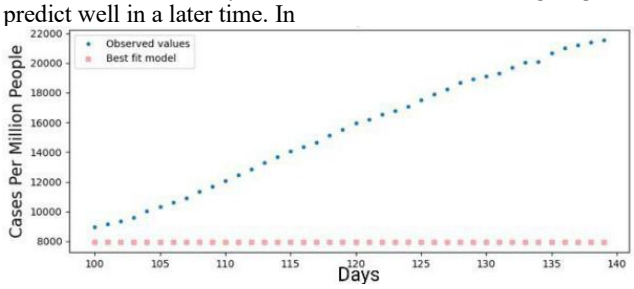

Figure 3. Regression results for logistic regression

C. *Results for RNN*

Since the emerging of deep learning, RNN gives a better choice for epidemic prediction. Compared with the traditional methods, such as linear regression and logistic regression, the performance is significantly superior to them. RNN can avoid the inaccuracy of linear regression and logistic regression because with complicated process, the predictive data fits the original really well. For our simple and preliminary trial, the MSE is about 2.8. Sometimes the MSE can even be close to zero. Though the performance is satisfying, it needs time and effort to predict essential data. The inference time is 3.89s.

IV. CONCLUSION

This paper aims to compare several traditional and deep-based models to predict the cases of COVID-19 in America. First, incomplete and noisy data are preprocessed. Then, the epidemic prediction is applied using those models. Finally, it can be concluded that the best model for pandemic prediction is RNN, with the lowest MSE being 2.8, which validates the effectiveness of the deep neural network. Furthermore, compared with traditional regression models, including linear regression and logistic regression, RNN achieves better results, but it still needs more depth exploration. This approach can help both the medical workers and the government to handle the COVID-19 issue and arrange supplements. Compared with traditional models like linear regression and logistic regression, the deep learning based models, for example, RNN, have strength in the overall accuracy, but it is inferior to inference time. Therefore, this paper suggests the method of how to select the proper models to predict the number of infected people. In the future, the researchers should focus on selecting more deep-based models, LSTM, and transformer network. Furthermore, it is necessary to construct more available datasets for future research and train deeper networks to exploit the advantage of RNN.

REFERENCES

[1] E. O. Nsoesie, R. Beckman, M. Marathe, and B. Lewis, “Prediction of an epidemic curve: A supervised classification approach,” Stat Commun Infect Dis., vol. 3, issue 1, 2011.

[2] P. Wang, X. Zheng, J. Li, and B. Zhu, “Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics,” Chaos, Solitons & Feactals, vol. 139, 2020.

[3] Y. Nan and Y. Gao, “A machine learning method to monitor China’s AIDS epidemics with data from Baidu trends,”. PLoS ONE, vol. 13, issue 7, 2018.

[4] D. A. Shah, E. D. De Wolf, P. A. Paul and L. V. Madden, “Predicting fusarium head blight epidemics with boosted regression Trees,” Phytopathology, vol. 104, issue 7, pp. 702-14, 2014.

[5] G. Pokharel and R. Deardon, “Supervised learning and prediction of spatial epidemics,” Spatial and Spatio-temporal Epidemiology, vol. 11, pp. 59-77, 2014.

[6] P. K. Attaluri, X. Zheng, Z. Chen, and G. Lu, “Applying machine learning techniques to classify H1N1 viral strains occurring in 2009 flu pandemic. Proceedings of 6th Annual Biotechnology and Bioinformatics Symposium, pp. 21-27, 2009 [BIOT-2009].

[7] N. Vida. Machine learning techniques to predict pandemic from social media. Master’s thesis, 2018.

[8] J. V. Kumar and S. Kumar, “An effective approach to track levels of influenza-A (H1N1) pandemic in India using twitter,” Procedia Computer Science, vol. 70, pp. 801-807, 2015.

[9] R. Singh, R. Singh, and A. Bhatia, “Sentiment analysis using Machine Learning technique to predict outbreaks and epidemics,” International Journal of Advanced Science and Research, vol. 3, issue 2, pp. 19-24, 2018.

[10] S. A. Shinde and P. R. Rajeswari, “Intelligent health risk prediction systems using machine learning: a review,” International Journal of Engineering & Technology, vol. 7, issue 3, pp. 1019-1023, 2018.

[11] K A Baharin, H A Rahman M Y Hassan , et al. Hourly irradiance forecasting for Peninsular Malaysia using dynamic neural network with preprocessed data[C]// Research & Development. IEEE, 2015.

[12] D T Larose, C D Larose. Discovering Knowledge in Data (An Introduction to Data Mining) || Data Preprocessing[J]. 2014, 10.1002/9781118874059:16-50.

[13] Zhang, Aoqian, Song, Shaoxu, Wang, Jianmin. Sequential Data Cleaning: A Statistical Approach.[J]. 2016.

[14] K J Preacher, P J Curran, D J Bauer Computational Tools for Probing Interactions in Multiple Linear Regression, Multilevel Modeling, and Latent Curve Analysis[J]. Journal of Educational and Behavioral Stats, 2006, 31(4):437-448.

[15] A. Ambainis, Polynomial degree vs. quantum query complexity[C]// es, 2006:220-238.

[16] Aluísio JD Barros, and V. N. Hirakata . "Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio." *BMC Medical Research Methodology* 3.1(2003):: 21.

[17] Liu S , Zhang S , Zhang X , et al. R-Trans: RNN Transformer Network for Chinese Machine Reading Comprehension[J]. IEEE Access, 2019:1-1.

[18] Qian J , Jiang L , Song Z . Locally linear back-propagation based contribution for nonlinear process fault diagnosis[J]. IEEE/CAA Journal of Automatica Sinica, 2020, 7(3):764-775.

[19] Rumelhart D E , Hinton G E , Williams R J . Learning Internal Representation by Back-Propagation Errors[J]. Nature, 1986, 323:533-536.

[20] Tallec C , Ollivier Y . Unbiasing Truncated Backpropagation Through Time[J]. 2017.