Epidemiological Investigation Model of Novel Coronavirus Pneumonia Based on Data Mining Technology

Kai Chen[1,2], Bing Yang[1], Miao Hao[1], Hong Yang, Meiyuan Qin[1], and Chengmei Zhang[1,*]

[1] Guizhou Academy of Sciences Big Data Co. LTD, Guiyang, Guizhou, 550000, China
[2] Guizhou Academy of Sciences, Guiyang, Guizhou, 550000, China

Abstract. With its extraordinary rapidity of transmission, the COVID-19 pandemic demonstrates the vulnerability of a globalized and networked world. The first months of the pandemic were marked by a significant strain on health-care systems. Since the prospect of pandemics has elevated public health concerns, it is critical to revisit this issue. The primary goal of this essay is to employ data mining technologies and methodologies to do investigative analysis on publicly available information. In this article we shared ways and techniques to handle and control this pandemic in the best possible way using data mining techniques and models. Researchers and scientists will be able to use the results of our poll to come up with new approaches to combat the pandemic.

1 Introduction

COVID-19 ("Corona"), a respiratory disease, was discovered to be spreading over the world in December 2019. The first confirmed case of infection was discovered in Wuhan, China. It started as an outbreak in China, but by the first half of 2020, it had turned into a pandemic that is still going on [1]. With its extraordinary rapidity of transmission, the COVID-19 pandemic demonstrates the vulnerability of a globalised and system or organization. The first months of the epidemic were marked by a significant strain on health-care systems. Severe limitations, such as current education system shutdowns, public transportation system failures, or a complete lockdown, were imposed on the people of countries around the world [11]. Many elements, such as the government, culture, and health system, influenced the intensity of the load. The burden, on the other hand, fell on each country with modest temporal gaps.

This research deals with data analytics for the COVID-19 pandemic's disease statistics. The purpose of this assessment is to evaluate/analyse infection data mining while taking into account measurement error, pandemic expanding behaviour with lockdown implications, and early second wave in several nations. The goal of our model is to predict the number of cases in the upcoming days. The data set included provides information about the number of deaths, confirmed cases, total recovered from death, and other details from a variety of nations where the pandemic wreaked havoc.

In this article, we used Data Mining technology [10], which consists of two technologies: one for preparing tasks that handle general data characteristics [14], and the other for creating predictive models, with the goal of building models that can estimate the mapping from inputs to outputs using a sample of data called training data.

2 Methods

In this section, we will go over the tactics and Data Mining techniques we used in our article to gain a better understanding of the data and improve the accuracy of our model for predicting Covid-19 instances in the future days.

2.1 Moving Average

A moving average is a statistical calculation that is used to analyse data points by calculating a series of averages of different subsets of the entire data set. A moving average (MA) is a stock indicator that is commonly used in technical analysis in finance [9]. The reason for using moving average in our model is check the number of cases, deaths and recovered people from the Covid-19 for the analysis purpose and, Its calculated with time gap of 7 days. Formula for calculating Simple Moving Average(SMA) is:

$$SMA = \frac{(W_1 + W_2 \pm - - - W_n)}{N} \quad (1)$$

- W is the average in period n
- N is the number of period n

2.2 Bayesian Ridge

Rather than using simple linear regression model we used Bayesian ridge model because it formulates linear regression using probability distributions rather than point

* Corresponding author: zcmei@gzbdi.com

estimates. The response, y, is not estimated as single value, but is assumed to be drawn from a probability distribution [13]. The model for Bayesian Linear Regression with the response sampled from a normal distribution is:

$$y \sim N(\beta^T X, \sigma^2 I) \quad (2)$$

The output, y, is derived from a normal (Gaussian) distribution with a mean and variance. The goal of Bayesian Linear Regression is to determine the posterior distribution for the model parameters rather than to find the single "best" value of the model parameters [2].

2.3 Support Vector Regressor

The second model we used for prediction is SVR, which works on the same principles as SVM for classification but with a few minor differences. In the case of regression, a margin of tolerance (epsilon) is set in order to approximate the SVM that the problem would have already requested. The formulation for SVR is [7]:

$$y = \sum_{i=1}^{N}(a_i - a_i^*) \cdot [\phi(x_i), \phi(x)] + b \quad (3)$$

3 Data Preprocessing and Analysis

Data set which we used consist of information about the Covid-19 cases globally. It contains data of total number of confirmed cases, deaths reported, recoveries and active cases all across the world. For analysis of data we used several techniques which are explained below.

3.1 Moving Average Analysis

We first used this to determine the total number of Covid-19 instances worldwide. We calculated the moving average over a 7-day period. As shown in Figure 1, the number of virus cases is increasing at an exponential rate every day. We can deduce from this that the epidemic will peak in the coming days.

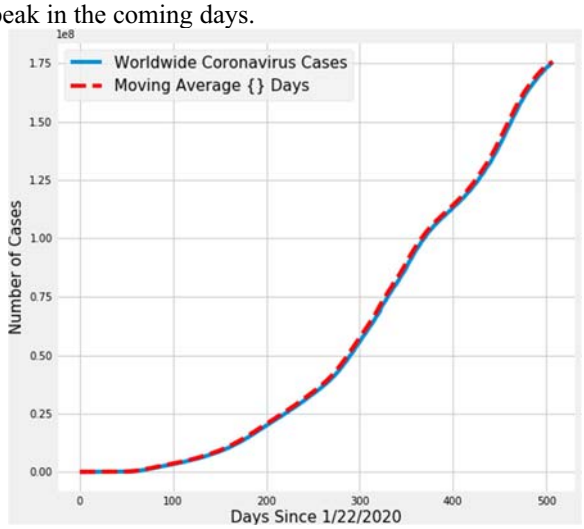

Figure 1. Worldwide Cases

After that, we utilised it to determine the global mortality toll. We can conclude from Figure 2 that the situation of Covid-19 is deteriorating day by day, as it is claiming the lives of a great number of people. For example, the death toll has surpassed 350000 after 500 days since January 2020, and it is continuing to rise at a rapid rate.

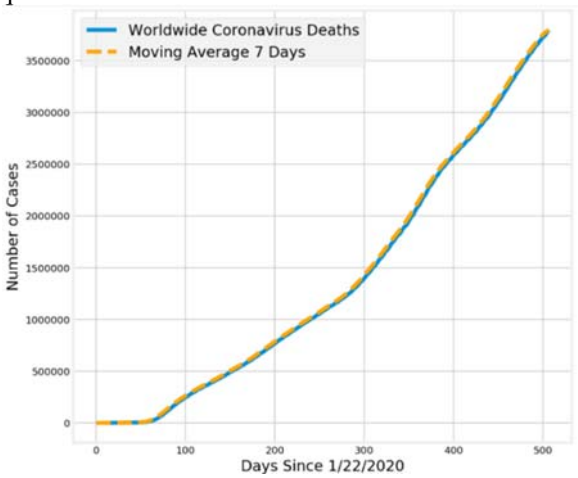

Figure 2. Death Toll

Then we calculated the number of cases that were increasing on a daily basis. Figure 3 shows that the daily increase in covid cases was quickly growing until 350 days after January 2020, However, after that, cases begin to decrease for a few days before increasing again, signalling to many countries that this could be the second wave of covid and alerting them to make health preparations accordingly.

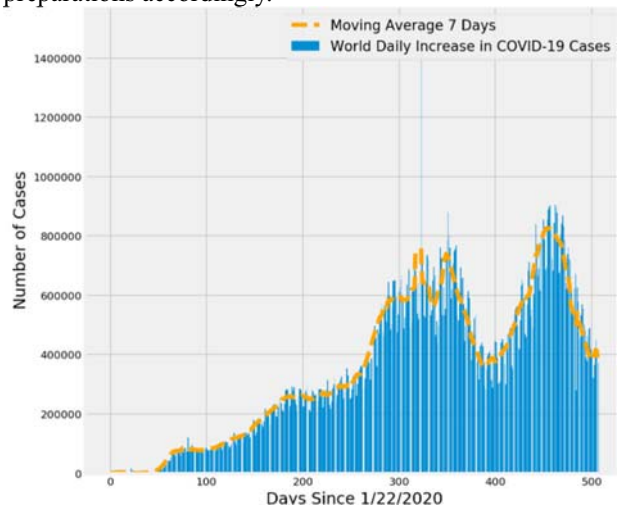

Figure 3. Daily Increase of Cases

3.2 Correlation Analysis

Correlation analysis is a statistical method for determining the strength of a relationship between two numerically measured continuous variables, such as a person's height and weight [3]. We used the following formula:

$$r = \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2(y_i - \bar{y})^2}} \quad (4)$$

Figure 5 shows that the correlation between confirmed cases and death is 0.933, implying that as the number of cases decreases, so does the death toll. However, the

correlation between recovered and active cases is 0.597, indicating that an increase in active cases has less of an impact on recovered cases [15].

	Confirmed	Deaths	Recovered	Active	Incident_Rate
Confirmed	1.000	0.931	0.913	0.971	0.198
Deaths	0.931	1.000	0.811	0.915	0.175
Recovered	0.913	0.811	1.000	0.791	0.240
Active	0.971	0.915	0.791	1.000	0.156
Incident_Rate	0.198	0.175	0.240	0.156	1.000

Figure 4. Correlation Table

4 Model

The complete system architecture of our model comprises of 3 models which are Deep Learning Model, Support Vector Regressor, and Bayesian Ridge Model as you can see it in Figure 5.

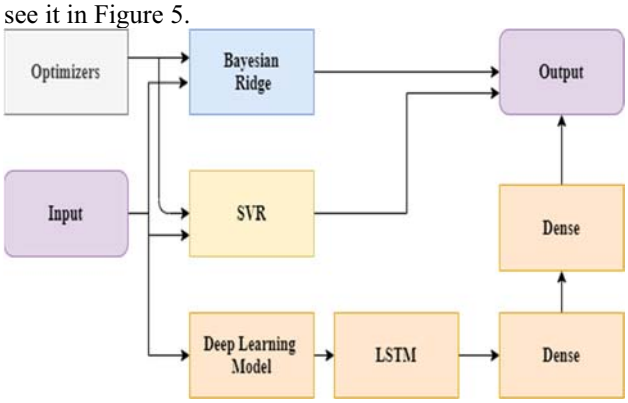

Figure 5. System Architecture

In the deep learning model first we passed the input to the input layer after that we applied LSTM layer [8] which is an artificial recurrent neural network [6] that is an effective model for the prediction of time series where data are sequential. By storing the past in hidden states, they can predict the outputs more accurately. In this study, the aim was to estimate the number of positive COVID-19 cases through time; as this is a well-suited task for the LSTM model, we used this model in our study. After adding LSTM layer we added 2 dense layers that is connected deeply, which means each neuron in the dense layer receives input from all neurons of its previous layer. The dense layer is found to be the most commonly used layer in the models. The output generated by the dense layer is an 'm' dimensional vector [12].

For the Bayesian ridge and SVR model we used hyper opt optimizers which is a way to search through hyper parameter space. For example, it can use the Tree-structured Parzen Estimator (TPE) algorithm [4], which explore intelligently the search space while narrowing down to the estimated best parameters. The parameters obtained for Bayesian ridge model are, alpha_1=1e-06, alpha_2=1e-06, compute_score=False, copy_X=True, fit_intercept=False, lambda_1=1e-06, lambda_2=1e-06, n_iter=300, normalize=False, tol=0.001.

5 Results

For the evaluation and performance of our models we used MSE as a loss function which shows how close a regression line is to a set of points [5]. From all of our model our Bayesian Regressor model performed best with MSE score of 0.89. As we can see in Figure 6 the predicted values Regressor line is very close to the actual value.

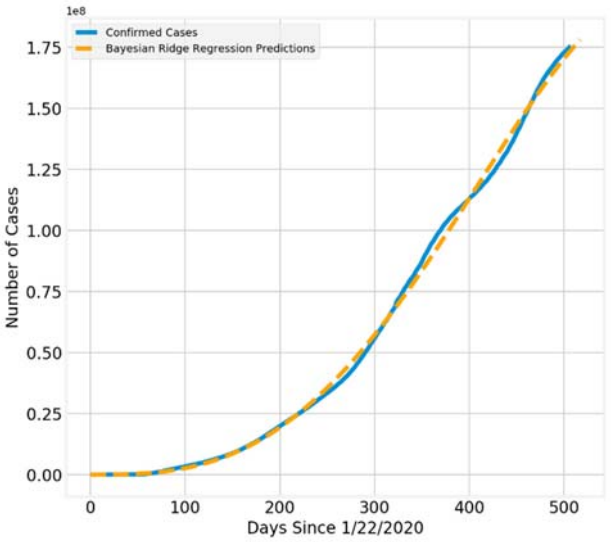

Figure 6. Bayesian Prediction

Second best RMSE was of SVR model with an score of 0.758.

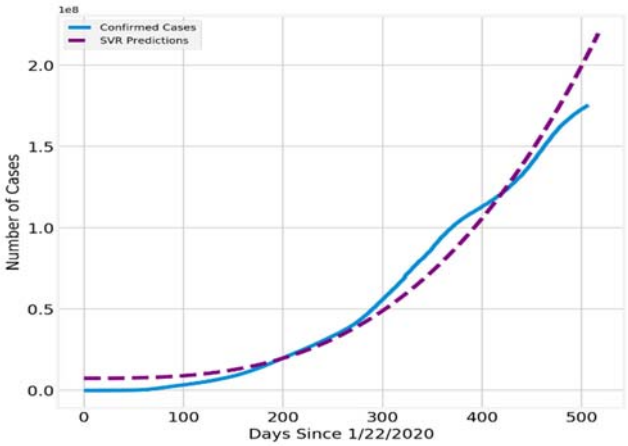

Figure 7. SVR prediction

6 Conclusion

The COVID-19 is one among the most deadly viruses that has greatly affected daily life affairs. The government and a number of other organizations should be interested to provide bases for comparison and to provide a better description of the data under consideration to get reliable estimates of the parameters of interest. In the present study, data mining models were developed for the prediction of COVID-19 infected patients' recovery using epidemiological dataset of COVID-19 patients. From all the models Bayesian Regression model performed best, SVR was on 2nd and deep learning model was giving

worst results and lthough there are a lot more to cover in the future.

Acknowledgement

This work was Supported by Project of Guizhou Academy of Sciences ([2020]11) and Guizhou Provincial Science and Technology Projects ([2020]4Y180）.

References

1. Thamina Acter, Nizam Uddin, Jagotamoy Das, Afroza Akhter, Tasrina Rabia Choudhury, and Sunghwan Kim. Evolution of severe acute respiratory syndrome coronavirus 2 (sars-cov-2) as coronavirus disease 2019 (covid-19) pandemic: A global health emergency. Science of the Total Environment, page 138996, 2020.
2. Douglas G Altman and J Martin Bland. Statistics notes: the normal distribution. Bmj, 310(6975):298, 1995.
3. Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In International conference on machine learning, pages 1247–1255. PMLR, 2013.
4. James Bergstra, Dan Yamins, David D Cox, et al. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In Proceedings of the 12th Python in science conference, volume 13, page 20. Citeseer, 2013.
5. Tianfeng Chai and Roland R Draxler. Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. Geoscientific model development, 7(3):1247– 1250, 2014.
6. Alexandre De Brébisson, Étienne Simon, Alex Auvolat, Pascal Vincent, and Yoshua Bengio. Artificial neural networks applied to taxi destination prediction. arXiv preprint arXiv:1508.00021, 2015.
7. Harris Drucker, Chris JC Burges, Linda Kaufman, Alex Smola, Vladimir Vapnik, et al. Support vector regression machines. Advances in neural information processing systems, 9:155–161, 1997.
8. Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. Neural computation, 12(10):2451–2471, 2000.
9. Gao-Feng Gu, Wei-Xing Zhou, et al. Detrending moving average algorithm for multifractals. Physical Review E, 82(1):011136, 2010.
10. Shu-Hsien Liao, Pei-Hui Chu, and Pei-Yuan Hsiao. Data mining techniques and applications–a decade review from 2000 to 2011. Expert systems with applications, 39(12):11303–11311, 2012.
11. Rizwan Rasheed, Asfra Rizwan, Hajra Javed, Faiza Sharif, and Asghar Zaidi. Socio-economic and environmental impacts of covid-19 pandemic in pakistan—an integrated analysis. Environmental Science and Pollution Research, 28(16):19926–19943, 2021.
12. Mazhar Shaikh, Ganesh Anand, Gagan Acharya, Abhijit Amrutkar, Varghese Alex, and Ganapathy Krishnamurthi. Brain tumor segmentation using dense fully convolutional neural network. In International MICCAI brainlesion workshop, pages 309–319. Springer, 2017.
13. Qi Shi, Mohamed Abdel-Aty, and Jaeyoung Lee. A bayesian ridge regression analysis of congestion's impact on urban expressway safety. Accident Analysis & Prevention, 88:124–137, 2016.
14. Zhifu Sun, Julie Cunningham, Susan Slager, and Jean-Pierre Kocher. Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. Epigenomics, 7(5):813–828, 2015.
15. Richard Taylor. Interpretation of the correlation coefficient: a basic review. Journal of diagnostic medical sonography, 6(1):35–39, 1990.