Data Prediction and Infection factor analysis in COVID-19

Xiaomei Ji

Department of Computer Science and Engineering, The University of New South Wales Sydney, Australia
xiaomei.ji@student.unsw.edu.au

Abstract
COVID-19 has become a world-wide pandemic since 2019. This paper focuses on data analysis in COVID-19 pandemic, which contains three main parts. First, the research aims to predict the daily death and cases using training data from single and multiple countries using four methods, including KNN, Decision tree, SVM and Linear regression. Then, the study further focuses on how the temperature will affect the COVID-19. Finally, the relationship of statistic between different countries also be discussed, such as daily increases and deaths. The linear regression achieves the best scores in daily statistic prediction. Furthermore, the experiment validates the positive relationship between the pandemic and temperature and also obtains that different countries has positive correlation with respect to the pandemic condition. The paper gives a throughout analysis on COVID-19 and exploits the essential factor of COVID-19 spread.

Keywords- *COVID-19; Daily statistic prediction; Factor Analysis; Machine learning.*

1. INTRODUCTION

Since COVID-19 broke out in 2020, according to the latest data, millions of cases confirmed the new crown pneumonia worldwide. There were 4696397 deaths worldwide. Data shows that in addition to China, countries with more than 10000 confirmed cases include Italy, the United States, Spain, Germany, Iran, France and other countries. The epidemic has spread to more than 200 countries and regions around the world.

Studies have shown that people aged 20-50 are more likely to be infected with COVID-19 by analyzing people of different age groups [1], and the time series prediction model FbProphet model based on machine learning can predict when the global epidemic will reach its peak [2].

There are many reasons for the outbreak and infection of COVID-19. When the population density in an area is high and the humidity is high, people in that area are more likely to be infected [3]. In addition to these two influencing factors, latitude may also be a cause affecting the transmission of COVID-19. Some surveys show that compared with countries in the northern hemisphere, the mortality in countries in the southern hemisphere is relatively low [4]. And when the climate begins to warm, the number of new cases will gradually decrease [5].

In this paper, we aim to focus on data prediction and exploit the factor of COVID-19, such as temperature and altitude, etc.

First, we conduct daily cases and deaths prediction using four different methods, including K-Nearest Neighbor (KNN), Supported Vector Machine (SVM), linear regression, decision tree. Then, we exploit the effectiveness of temperature in COVID-19. Finally, we introduce the relationship of pandemic condition between different countries, including South-Korea, India, Italy.

Above all, we conclude that our models are capable of pandemic prediction. Specifically, linear regression achieves the best performance in most of prediction cases. As for the daily cases' prediction in South-Korea, SVM achieves the best result and when predicting the daily death for Italy, decision tree gets the top performance. The linear regression also achieves the best performance when training using multiple countries, with 0.874 of R2 for daily cases and 0.729 for daily death. It can be seen from the experiment that the daily cases / daily death of each country is related to the local temperature, which may be positive or negative. And daily death / daily cases in different countries are also related. We propose an in-depth research on different aspects on COVID-19 analysis

S. A. R. Khan et al. (Eds.): ICMSSE 2022, AHCS 12, pp. 226–231, 2023.
https://doi.org/10.2991/978-94-6463-056-5_32

2. METHOD

2.1.Data acquisition

In terms of data acquisition, we first browsed the data provided by various projects on COVID-19 on kaggle (https://www.kaggle.com/aestheteaman01/covcsd-covid19-countries-statistical-dataset), and finally found a data package containing dozens of countries. Through screening the data of all countries, we finally selected the data of three specific countries (India, South Korea and Italy) for subsequent experiments.

The data content of each country includes 25 contents, including cumulative_cases, cumulative_death, daily_cases, daily_death, temperature, wind_speed, population, confirmed_cases, life expectancy and so on.

In terms of data preprocessing, firstly, some unnecessary features are deleted. Secondly, it is found that the data amount of some features is very various. If the original index value is directly used for analysis, the role of the index with higher value in the comprehensive analysis will be highlighted and the role of the index with lower value level will be relatively weakened. Therefore, feature normalization is utilized for an accurate prediction.

2.2.Data prediction using four methods in a single country as well as multiple countries

In this experiment, four models were used to predict the daily growth and death of patients in each country through other features. After that, the data of the three countries were combined, and the combined data were used to predict the daily growth of growth and death of patients. These four models are decision tree, SVM, KNN and linear regression.

Decision tree is a prediction model of attribute structure, which represents a mapping relationship between object attributes and object values. It consists of nodes and directed edges. There are two types of nodes: internal nodes and leaf nodes. Internal nodes represent a feature or attribute, and leaf nodes represent a class. In essence, the learning of decision tree is to summarize a set of classification rules from the training set to obtain a decision tree with less contradiction with the data set and has good generalization ability. The loss function of decision tree is usually a regularized maximum likelihood function. Heuristic methods are usually used to approximately solve this optimization problem. Decision tree learning steps include feature selection, decision tree generation and decision tree pruning. Decision trees represent a conditional probability distribution, so different depth of decision trees corresponds to different complexity. The generation of decision tree corresponds to the local selection (Local Optimization) of the model, and the pruning of decision tree corresponds to the global selection (Global Optimization).

SVM is a binary classification model. Its basic model is the linear classifier with the largest interval defined in the feature space, which makes it different from the perceptron; SVM also includes kernel techniques, which makes it a substantially nonlinear classifier. The learning strategy of SVM is interval maximization, which can be formalized as a problem of solving convex quadratic programming, which is also equivalent to the minimization of regularized hinge loss function. The learning algorithm of SVM is the optimization algorithm for solving convex quadratic programming.

KNN is a machine learning algorithm that can be used for classification and regression. For a given test sample, find the K training samples closest to it in the training set based on the distance measurement, and then predict based on the information of the K "neighbors".

Linear regression is a kind of regression analysis, which indicates that there is a linear relationship between independent variables and dependent variables. Regression analysis starts from data, investigates the quantitative relationship between variables, describes this relationship through a certain mathematical relationship, estimates the value of a variable through the relationship, and gives the reliability of the estimation.

In our experiments, we take the previous 59 days for training those models and use the last 10 days for evaluation. Furthermore, in training process using the data from multiple countries, we firstly concatenate those training data from three countries as well as the test data. In evaluation step, we predict those data together.

2.3.Exploit the relationship between pandemic condition and temperature.

In this experiment, Pearson correlation coefficient is mainly used for analysis.

Pearson correlation coefficient, also known as Pearson product moment correlation coefficient, is a linear correlation coefficient and the most commonly used correlation coefficient. It is recorded as R to reflect the linear correlation between the two variables X and Y. The value of R is between - 1 and 1. The greater the absolute value, the stronger the correlation.

Pearson correlation coefficient reflects the strength of the linear correlation between the two variables. The greater the absolute value of R, the stronger the correlation.

When $R > 0$, it indicates that the two variables are positively correlated, that is, the greater the value of one variable, the greater the value of the other variable;

When R < 0, it indicates that the two variables are negatively correlated, that is, the larger the value of one variable, the smaller the value of the other variable;

When R = 0, it indicates that the two variables are not linearly correlated (note that it is only nonlinear correlation), but there may be correlation in other ways (such as curve mode);

When r = 1 and - 1, it means that the two variables X and y can be well described by the linear equation, and all sample points fall on a straight line.

Pearson correlation coefficient can be calculated in three forms as follows:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}} \quad (1)$$

$$\rho_{X,Y} \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)\hat{}2}{N})(\sum Y^2 - \frac{(\sum Y)\hat{}2}{N})}} \quad (2)$$

$$\rho_{X,Y} \frac{N\sum XY - \sum X \sum Y}{\sqrt{(N\sum X^2 - (\sum X)\hat{}2)}\sqrt{N\sum Y^2 - (\sum Y)\hat{}2}} \quad (3)$$

The experimental method is to compare the daily death / cases of different countries with min_ temperature, max_ temperature, avg_ Temperature calculates the Pearson coefficient. Thus, the correlation between the number of new patients per day, the number of dead patients per day and temperature is determined.

2.4.Exploit the relationship of pandemic condition among countries

The above experiments are data analysis of a single country. Then we combine the data of the three selected countries, and analyze the relationship between them by calculating the Pearson coefficient described above between daily death and daily cases.

2.5.Implementation details

Our method is implemented using Python and scikit-learn package. As for linear regression, we set the parameter 'fit_intercept' to True, 'Normalize' to False, 'copy_X' toTrue, and 'n_jobs' to None. In SVM, we apply an non-linear kernel, which is denoted as 'rbf'

3. RESULTS AND DISCUSSION

3.1.Prediction performance using single country and multiple countries

In this section, we aim to evaluate those four models using R^2 as our evaluation score. The value range of R^2 is R^2<=1, the larger it is, the more accurate the prediction result is.

Table 1 and Table 2 show the prediction results of daily cases and daily death using the four models respectively. The score shows the ability of the model to predict the results.

It can be seen from the scores that SVM is the best model for South-Korea, linear is the best model for India, and linear is the best model for Italy.

At the same time, we find that there are various reasons affecting the prediction results, which may be the setting of model parameters or the amount of data.

Since we use the data of the first 59 days to predict the data of the last 10 days, the amount of data may be the main factor affecting the results, and we found that in the data of the first 59 days, the data of the first few days basically did not change or even 0, which also had some negative effects on the prediction of the model.

TABLE 1. THE DAILY CASES SCORE OF SINGLE COUNTRY BY USING FOUR MODELS.

	KNN	SVM	Linear	Decision Tree
South-Korea	-0.37	-0.03	-283.80	-0.77
India	-1.81	-2.00	0.02	-1.00
Italy	-11.75	-8.66	-0.06	-0.74

TABLE 2. THE DAILY DEATH SCORE OF SINGLE COUNTRY BY USING FOUR MODELS.

	KNN	SVM	Linear	Decision Tree
South-Korea	-1.10	-1.39	0.06	-2.06
India	-0.73	-0.69	0.09	-0.45
Italy	-21.57	-21.52	-1.96	-1.82

Figure 1 and Figure 2 show the prediction after the data of three countries are combined. Figure 1 represents the prediction result of daily case. Figure 2 represents the prediction result of daily death. After data combination, we can clearly see that the prediction results are much more accurate than those of a single country, which also proves that the size of the data does affect the prediction results of the model.

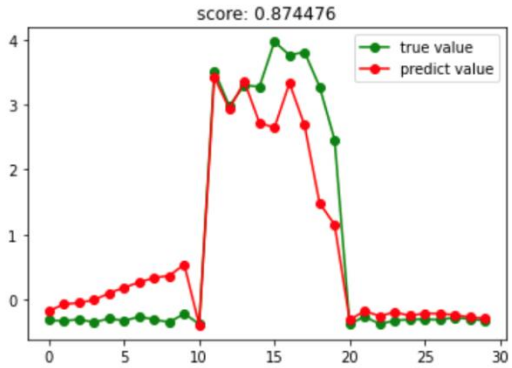

Figure.1 The visualization of predicted daily cases for all three countries. We conclude that the linear regression model achieves an accurate performance.

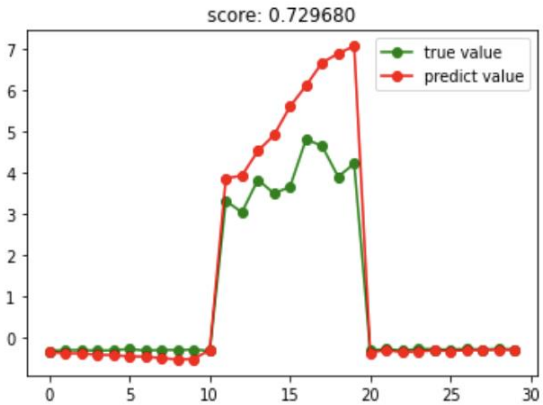

Figure.2 The visualization of predicted daily deaths for all three countries. We conclude that the linear regression model achieves an accurate performance.

3.2. Temperature analysis on COVID-19

Table 3,4,5 show the correlation between daily death/daily cases of each country and temperature, min_temperature, max_temperature respectively.

Through the calculated Pearson coefficient, we can see that daily death / daily cases is related to temperature. And South-Korea, India, Italy are positively correlated with temperature, that is, the higher the temperature, the more the number of confirmed cases and deaths. However, daily death / daily cases in Italy are negatively correlated with min_temperature, that is, the higher the minimum temperature of the day, the less the number of confirmed cases and deaths.

TABLE 3. THE RELATIONSHIP BETWEEN THE PANDEMIC CONDITION AND TEMPERATURE IN SOUTH KOREA.

Correlation	Temperature	Min_temperature	Max_temperature
Daily_death	0.49	0.22	0.55
Daily_cases	0.11	0.12	0.14

TABLE 4. THE RELATIONSHIP BETWEEN THE PANDEMIC CONDITION AND TEMPERATURE IN INDIA.

Correlation	Temperature	Min_temperature	Max_temperature
Daily_death	0.31	0.40	0.30
Daily_cases	0.54	0.58	0.52

TABLE 5. THE RELATIONSHIP BETWEEN THE PANDEMIC CONDITION AND TEMPERATURE IN ITALY.

Correlation	Temperature	Min_temperature	Max_temperature
Daily_death	0.29	-0.11	0.32
Daily_cases	0.37	-0.08	0.38

3.3.Relationship of pandemic condition between different countries

TABLE 6. THE RELATIONSHIP AMONG COUNTRIES FOR DAILY CASES DATA. WE REMOVE THE SELF-CORRELATION TO CLARIFY THE RESULTS.

	India	Italy	South-Korea
India	-	0.63	0.40
Italy	0.63	-	0.62
South-Korea	0.40	0.62	-

TABLE 7. THE RELATIONSHIP AMONG COUNTRIES FOR DAILY DEATH DATA. WE REMOVE THE SELF-CORRELATION TO CLARIFY THE RESULTS.

	India	Italy	South-Korea
India	-	0.79	-0.05
Italy	0.79	-	-0.03
South-Korea	-0.06	-0.03	-

Table 6 shows whether there is a correlation between daily cases in three countries. Table 7 shows whether there is a correlation between daily death in three countries.

For daily cases, any two of the three countries are positively correlated. For daily death, only India and Italy are positively correlated.

4. CONCLUSION

In this paper, we aim to exploit the data analysis on COVID-19 from three aspects, First, we apply four different methods to predict the trend of COVID-19 statistic. Our linear regression model achieves 0.874 of R^2, which indicates an accurate performance. Then, we also conduct on how the temperature will influence the trend of COVID-19. Finally, we obtain the correlation of pandemic condition among three countries. It can be seen from the experiment that the daily cases / daily death of each country is related to the local temperature, which may be positive or negative. And daily death / daily cases in different countries are also positively related.

In the future, in order to make the prediction results more accurate, we can utilize more powerful neural networks, such as CNN and RNN, to achieve an accurate prediction or integrate the data of more countries and find

more features that may affect the prediction results for training.

REFERENCES

[1] K. B. Prakash, S. S. Imambi, M. Ismail, T. P. Kumar, and Y. N. Pawan, "Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms," International Journal of Emerging Trends in Engineering Research, vol. 8, no. 5, pp. 2199-2204, May 2020.

[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 & Fractals, vol. 139, October 2020.

[3] M. Ahmadi, A. Sharifi, S. Dorosti, S. J. Ghoushchi, and N. Ghanbari, "Investigation of effective climatology parameters on COVID-19 outbreak in Iran," Science of The Total Environment, vol. 729, August 2020.

[4] J. M. Rhodes, S. Subramanian, E. Laird, and R. A. Kenny, "Editorial: low population mortality from COVID-19 in countries south of latitude 35 degrees North supports vitamin D as a factor determining severity," PMC, April 2020.

[5] A. Notari, "Temperature dependence of COVID-19 transmission," Science of The Total Environment, vol. 763, April 2021.