บทความวิจัย (Research Article)

Predictive Analysis of COVID-19 Patients in Thailand using Multiple Countries Data

Siratee Vorathamthongdee, Prabhas Chongstitvatana*

Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University Bangkok, Thailand
* Corresponding author
prabhas.c@chula.ac.th

Received: 26 Feb 2023
Revised: 19 Apr 2023
Accepted: 21 Apr 2023

Abstract

COVID-19 is a situation that has spread worldwide since 2019. This study predicts the number of patients with COVID-19 in Thailand. Using data between January 22, 2020, and December 31, 2021, we collect confirmed cases from John Hopkins open data. Prediction of the number of patient cases in the country helps the government manage its policies and resources. In this study, the K-Means clustering algorithm is performed to group the countries that have similar patterns of confirmed cases to Thailand. Clustering results show that Japan, Malaysia, Philippines, Bangladesh, Cuba, Iraq, Mexico, and Vietnam are all in the same cluster as Thailand. Long Short-Term Memory (LSTM) is used to predict confirmed cases in Thailand. The training data consists of the data from the pair of countries in the same cluster as Thailand. The result shows that pairing the data of Bangladesh, Japan, and Mexico with Thailand has the lower error on MAPE than using only Thailand data.

Keywords: COVID-19, Machine Learning, K-Means, LSTM

1. Introduction

The World Health Organization (WHO) reports the world first viral infection in Wuhan, China. December 31, 2019, marks the beginning of the coronavirus disease starting in 2019, also known by the abbreviation COVID-19. The infection can cause symptoms such as fever or a respiratory disorder. In severe cases, it can invade the lungs and destroy tissue until the patient is unable to breathe and eventually dies. The virus can spread quickly and show no symptoms [1]. In 2020, January WHO publicly announced the guidelines for government authorities, health workers, and other key stakeholders to guide response to community spread [2]. In the same month, there is a patient infected with COVID-19 in Thailand, which is the first patient infected outside of China [3]. The epidemic was spreading quickly, and it was later reported by the WHO on January 30, 2020, that the virus had spread to 18 countries around the world. On March

11, the number of cases had reached more than 100,000 in 114 countries around the world. It was a crisis that governments around the world had to face.

Cloud computing and artificial intelligence (AI) are important factors that influence the success of healthcare technology. Today, data is generated as a result of smart devices, which has encouraged the role of machine learning in this industry [4]. Daily analysis and forecasting of the number of COVID-19 cases is very beneficial to the government in anticipation of the epidemic. The prediction data helps the government plans public health policy to cope with patients.

In anticipation of the epidemic, the government benefits greatly from daily analysis and forecasting of the number of COVID-19 cases. The prediction data supports policy planning, personnel management in both public and private sectors, hospital bed allocation, and medical equipment allocation such as breathing apparatus. These forecasts also help in strategizing vaccine importation or production to meet the population's needs, as well as in ensuring a steady supply of essential medicines as the number of patients continues to rise. Forecasting directly benefits the economy and people, such as the tourism sector, which accounted for 11% of Thailand's GDP in 2019 before the COVID-19 outbreak [5], as well as a new way of living that is adapted to the situation.

The analysis of COVID-19 cases is characterized by the data on the daily increase in the number of cases. The approach to forecasting the number of patients involves various machine learning techniques to discover the pattern of epidemic data. Various techniques have been used, such as Support Vector Regression [6], Deep Learning models including Recurrent Neural Network (RNN), [7] Gated Recurrent Unit (GRU) [8], Long Short-Term Memory (LSTM) [9], and others. The goal of this study is to use machine learning for analyzing and predicting the number of COVID-19 patients.

COVID-19 patient data from Johns Hopkins University is used in our study. There have been 2 years of times series of confirm cases following the COVID case. Multiple data can improve prediction accuracy. Therefore, we aim to use data from multiple countries to predict the case in Thailand. Using the k-means algorithm, all country data are clustered into groups, and only countries that are similar to Thailand are chosen. We use LSTM to predict confirmed cases of COVID-19 in Thailand using country data from the same Thailand group. Finally, we provide data analysis and visualization of the outcome of the confirmed case.

The remainder of this paper is structured as follows. Section 2 contains related works. Section 3 contains details about the research methodology. The experiment results and discussion are described in Section 4. Section 5 provides the conclusion.

2. Related Work

Many studies have been conducted to predict the number of COVID-19 cases, for example, [10] use K-means clustering to cluster data from 155 countries into groups, using attributes such as population, air quality index and health system indicators. The work in [11] determines the COVID-19 cases in Indonesia by grouping provinces using the similarity of data of each province based on confirmed cases, recovered cases and death cases.

Several studies used LSTM to make predictions, for example, [12] uses LSTM model to study the spread of virus in Canada; [13] uses LSTM and GRU to predict the death cases of COVID-19; [14] uses LSTM to predict

the recovered cases, daily positive cases and deceased cases in India; [15] performs forecasting in six countries using ARIMA, LSTM, and Transformer; [16] proposes COMAP using clustering and prediction to predict COVID-19 cases; [17] uses K-means to cluster group of countries with similar indicators, then uses LSTM, ARIMA, SMA-6, D-EXP-MA, to predict the number of cases.

3. Method

3.1 Data Processing

Our study use data from Johns Hopkins Center for Systems Science and Engineering open data on Github which is updated daily. Confirmed cases are collected for all countries between January 22, 2020, and December 31, 2021. The total number of data points is 710. We keep only the confirm case and remove all non-related features from this study. With a total of 199 countries for all data, we discovered eight countries with keep-in-region levels, namely Australia, Canada, China, Denmark, France, the Netherlands, New Zealand, and the United Kingdom, as well as other countries with keep-in-country levels. The data from those 8 countries are aggregated into countries level. Some missing data or null values are replaced with zero values. After that, the cumulative confirmed cases are converted to non-cumulative cases to use them as real daily confirmation cases. The data is normalized into a 0-1 range.

3.2 Clustering Country

In order to select the number of clusters, we use the Elbow curve method. This method determines the optimal K value by averaging the inertia, the sum of distance values between the centroid point and every point starting from 1 to the desired n. The appropriate K is determined by the value located at the point of the angle, like an elbow, by selecting the largest error difference compared to the previous K value. The number of clustering from 2 to 50 clusters are used to compare the Inertia measure.

K-Means clustering [18] is one of the unsupervised learning algorithms of data mining. The fundamental idea is to set the cluster's center (Centroid) to k points, define the centroid in each data point randomly, calculate the distance between the datasets and the center using Euclidean distance, and then cluster the dataset with a minimal distance. The centroids are recalculated in each cluster by computing the Euclidean distance, and this process is repeated until the centroid points no longer change.

To improve the standard K-Means algorithm, we also apply the K-Means++ algorithm [19], which enhances the initial selection of centroids. Standard K-Means has the disadvantage of being sensitive to the initial centroid values, which can affect the distance calculation between centroids and data points, ultimately impacting the subsequent clustering. K-Mean++ starts by randomly selecting the first centroid on the data point, and then calculating the distance between all data points and centroids according to the equation.

$$d_i = \max_{j:1\rightarrow m} \|x_i - c_j\|^2 \quad (1)$$

Equation 1 Equation to choose centroid of k-mean++

The equation 1 show that d_i is the distance between farthest data point (x_i) and centroid (c_j) where m represents the number of selected centroids. Next step, select the data point (x_i) to be a new centroid by selecting from the d_i value or the point that is the most distant from the centroid. Then repeat until get a new centroid that has been set completely. When initial

centroid is finished, after that, do the same as standard K-Means.

3.3 Alignment of Pair Countries

To measure the similarity between countries in the process of grouping, we consider time-shift properties. Time-shift is the observation that when compare the number of cases between two countries, the data may be similar but there is time difference between them. Therefore, when measuring the similarity of two data, we fix one data and perform time-shift of other data and find the maximum correlation. In the experiment, Thailand data is fixed, and data of other countries are shifted. We fill the empty data after shifting with the extrapolate value. The other pre-processing is the smoothing by using a moving average from 7 days and normalization, before doing the time-shift.

We divide training and testing data in a 90:10 ratio. The training data ranges from January 23, 2020, to October 22, 2021, and we use 10% of the training data for validation. Testing data has ranged from October 23, 2021, to December 31, 2021. We train data on multiple lag timestep using data from the previous N-th time step and predicting on timestep N+1. The window size is 14 days. We use the data from the countries that are in the same cluster of Thailand. The data of a pair of a country with Thailand is used as training data for the LSTM model to predict confirmed Thailand cases.

3.4 Forecast COVID-19 Cases

Long Short-Term Memory [20] is one type of Recurrent Neural Network (RNN). It has an internal gate that prevents the vanishing and exploding gradients of RNNs.

We use multiple layers of LSTM. The input layer is (14,2) which is 14 days data and 2 for the pair of a country with Thailand data. The first layer has 64 cells with an input timestep. Using dropout 0.2 to regularize to reduce the problem of overfitting. We then use an LSTM layer with 32 cells which are half of the first layer. The final layer is the fully connected node of the output result (Fig.1).

Layer (type)	Output Shape	Param #
lstm_246 (LSTM)	(None, 14, 64)	17152
dropout_123 (Dropout)	(None, 14, 64)	0
lstm_247 (LSTM)	(None, 32)	12416
dense_123 (Dense)	(None, 1)	33

Fig. 1 LSTM Structure Model

We use a learning rate of 0.001 with a reduced learning rate on Plateau. This method reduces the learning rate when the metrics are not improving. If the metrics stop improving after 5 epochs, we reduce the learning rate by multiplying by 0.2. The maximum epoch round is 50. We report the best epoch round on the best validation loss and use it for prediction. All experiments are running 30 times. Finally, we evaluate the performance and find the average of each metric.

3.5 Evaluation Metrics

To assess the performance of LSTM predictions, we use the root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) when comparing the forecasted results against actual values.

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{Actual_i} - \mathrm{Predicted_i})^2} \quad (2)$$

Equation 2 Equation of root mean square error (RMSE)

$$\mathrm{MAE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}|\mathrm{Actual_i} - \mathrm{Predicted_i}| \quad (3)$$

Equation 3 Equation of mean absolute error (MAE)

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\text{Actual}_i - \text{Predicted}_i}{\text{Actual}_i}\right| \text{x } 100 \quad (4)$$

Equation 4 Equation of mean absolute percentage error (MAPE)

where Actual_i= real COVID-19 data

Predicted_i= predicted COVID-19 data from model

The RMSE is the square root of the mean of the squared differences between the predicted value from model and actual testing data. The MAE is defined as a measure of the average absolute difference between the predicted values and their corresponding actual values. The MAPE measures the average percentage difference between the predicted values and the actual values. It is calculated by taking the absolute difference between each predicted value and its corresponding actual value, dividing it by the actual value, and then taking the average of the resulting percentages. All the metrics RMSE, MAE, and MAPE are expressed such that a lower value indicates better performance. During model training, we use the MAE as the loss function. To evaluate the performance, we use RMSE, MAE, and MAPE metrics.

4. Result and Discussion

4.1 Clustering Country

Firstly, we select the number of clusters as 8 from Fig. 2 because it is the first elbow, but the results after doing the K-Means show that Thailand is in the biggest cluster and there are 183 countries in the same cluster as Thailand. So we change to select the number of next elbows with the number of clusters of 15 as shown Fig. 3. After applying the K-Mean cluster algorithm to group countries into groups. The results from doing K-Means will get all 8 countries in the same cluster as Thailand, namely Japan, Malaysia, Philippines, Bangladesh, Cuba, Iraq, Mexico, and Vietnam, as shown in Fig. 4.

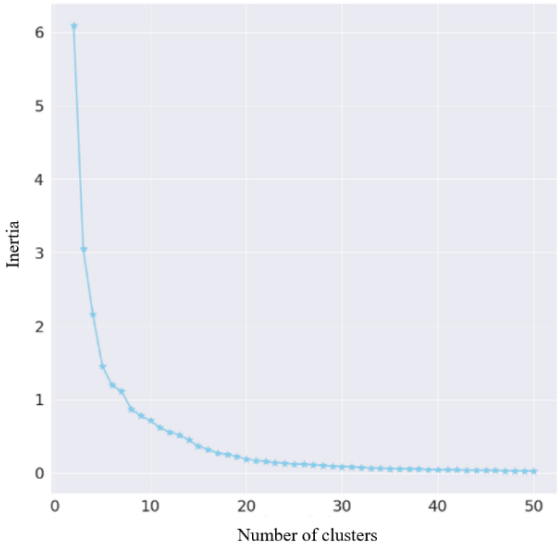

Fig. 2 Inertia of K-Means number

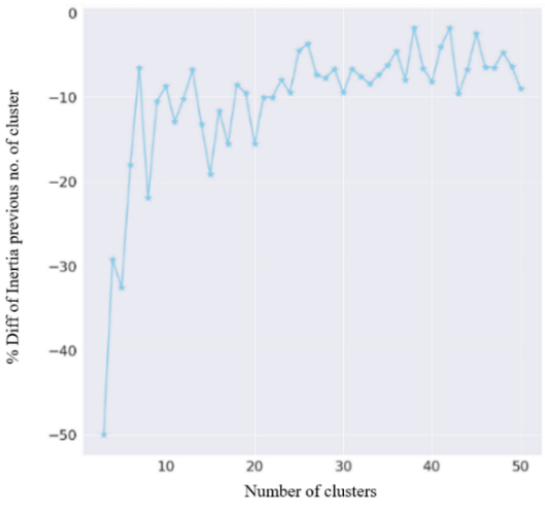

Fig. 3 Percent difference inertia previous number of clusters between 2 and 50

4.2 Alignment Result

Upon obtaining the result of K-Means algorithm, countries within the same cluster will be paired with Thailand, followed by an alignment process. The

Pearson correlation is calculated to determine the position with the highest correlation value for shifting the paired country's data to that position.

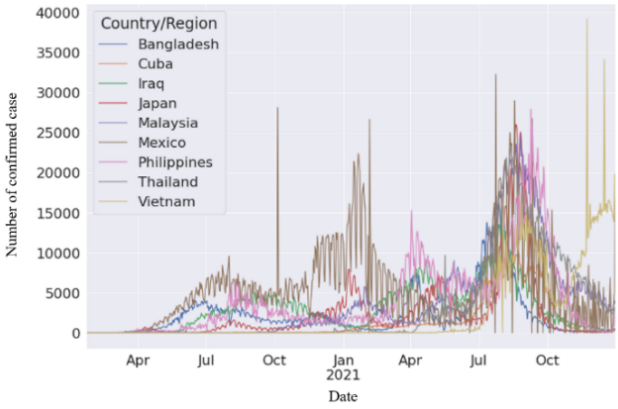

Fig. 4 All countries that are in the same cluster as Thailand.

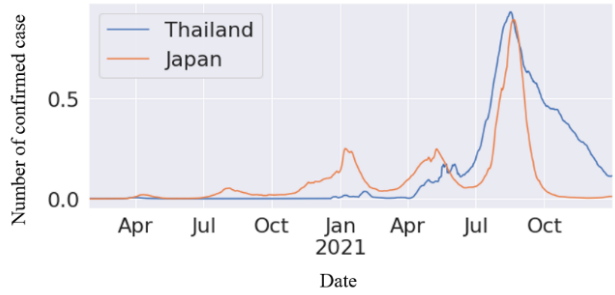

Fig. 5 Data alignment between Thailand and its cluster (Example for Japan and Thailand).

An example of data alignment is shown in Fig. 5. After normalizing and smoothing the data of Thailand and Japan, the data shows Japan data needs to be shifted (to the left) 4 days to have the highest correlation value.

The result in Table 1 demonstrates that following the alignment process, all countries display higher correlation values than before. The countries that show the highest daily shifting values are Iraq, Bangladesh, and Vietnam, in that order. Conversely, the countries with the lowest shifting values are Malaysia, Japan, and Mexico, respectively.

Table 1 Comparing correlation before and after processing data alignment between Thailand and its cluster

Country	Correlation Before shifting	Correlation After shifting	Best Lag
Japan	0.71	0.73	-4
Malaysia	0.95	0.96	-1
Philippines	0.70	0.75	-15
Bangladesh	0.50	0.75	29
Cuba	0.92	0.94	8
Iraq	0.51	0.70	30
Mexico	0.45	0.55	6
Vietnam	0.64	0.69	-24

4.3 Forecasting Result

Table 2 shows all results of LSTM from every pair of countries and gets an average of 30 experiments of all metrics and calculates the independent T-Test between only using Thailand data and all other pair countries. The result with a star on the number shows that the mean value is significantly different at a 0.05 level of significance when compared to using only Thailand's data for the same metrics.

The LSTM model results indicate that incorporating the data of the country that paired with Thailand improve the accuracy of the prediction in three metrics RMSE, MAE, and MAPE. It leads to better performance compared to using Thailand data alone. Japan, Philippines, Bangladesh, Iraq, and Mexico all have lower error values when paired with Thailand. Malaysia has a lower MAPE value than Thailand. Cuba

and Vietnam show higher error values across all three metrics compared to Thailand alone.

Table 2 Result from average 30 experiment of metric RMSE, MAE and MAPE from LSTM and T-Test using Thailand data and all other pair

Country	RMSE	SD	MAE	SD	MAPE	SD
Thailand	919.48	256.24	786.46	252.39	17.21	5.49
Thailand + Japan	762.85*	255.19	633.87*	253.03	13.23*	5.68
Thailand + Malaysia	973.48	358.44	792.01	316.30	13.89*	4.91
Thailand + Philippines	909.05	307.75	758.50	271.30	13.72*	4.65
Thailand + Bangladesh	822.35	206.60	685.79	196.60	12.86*	3.23
Thailand + Cuba	1426.10*	418.05	1257.39*	414.96	22.59*	7.80
Thailand + Iraq	863.10	291.41	732.56	273.09	13.66*	4.68
Thailand + Mexico	782.92	281.91	655.90	277.51	13.34*	6.00
Thailand + Vietnam	3273.73*	790.79	3086.91*	780.92	73.27*	17.95

MAPE of Japan, Malaysia, Philippines, Bangladesh, Iraq, and Mexico is less than using Thailand alone. When the result of pair countries are compared each other, it is found that there is no significant difference.

John Hopkins data source contains information on COVID-19 patients without separating each species of the disease. Data may be limited in terms of different strains of the disease. A development approach in terms of information can be done by finding a source of information by adding information about the species of the infected person to make the model able to learn the behavior and differences of each species. The size of the data set over 2 years is not enough. The limited amount of data can be solved by using techniques such as data augmentation.

5. Conclusion

We use K-means clustering to find the group of countries that have the same pattern of confirmed COVID-19 cases. The countries that are in the same group of Thailand are selected and their data are used in conjunction with data from Thailand. LSTM is used as machine learning method. The analysis shows that the best result (having lowest MAPE) is achieved when Thailand data is augmented with data from Bangladesh, Japan, and Mexico.

6. References

[1] O. L. Aiyegbusi et al., "Symptoms, complications and management of long COVID: a review," Journal of the Royal Society of Medicine, vol. 114, no. 9, pp. 428-442, 2021.

[2] W. H. Organization, "Responding to community spread of COVID-19: interim guidance, 7 March 2020," World Health Organization, 2020.

[3] WHO. (7 June 2022). *Archived: WHO Timeline - COVID-19*. [Online] Available: https://www.who.int/news/item/ 27-04-2020-who-timeline---covid-19

[4] T. Panch, H. Mattie, and L. A. Celi, "The "inconvenient truth" about AI in healthcare," NPJ digital medicine, vol. 2, no. 1, pp. 1-3, 2019.

[5] P. Warittha, Suchanan, C. (10 June 2022). *Tourism at a crossroad*. Bank of Thailand. [Online] Available: https://www.bot.or.th/Thai/ResearchAndPublications/articles/Pages/Article_18Aug2021.aspx
[6] T. Mantoro, R. T. Handayanto, M. A. Ayu, and J. Asian, "Prediction of covid-19 spreading using support vector regression and susceptible infectious recovered model," in 2020 6th International Conference on Computing Engineering and Design (ICCED), 2020: IEEE, pp. 1-5.
[7] R. L. Kumar, F. Khan, S. Din, S. S. Band, A. Mosavi, and E. Ibeke, "Recurrent neural network and reinforcement learning model for COVID-19 prediction," Frontiers in public health, vol. 9, 2021.
[8] L. Bi, M. Fili, and G. Hu, "COVID-19 forecasting and intervention planning using gated recurrent unit and evolutionary algorithm," Neural Computing and Applications, pp. 1-19, 2022.
[9] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," Physica D: Nonlinear Phenomena, vol. 404, p. 132306, 2020.
[10] R. M. Carrillo-Larco and M. Castillo-Cara, "Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach," Wellcome open research, vol. 5, 2020.
[11] D. Abdullah, S. Susilo, A. S. Ahmar, R. Rusli, and R. Hidayat, "The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data," Quality & Quantity, vol. 56, no. 3, pp. 1283-1291, 2022.
[12] V. K. R. Chimmula and L. Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks," Chaos, Solitons & Fractals, vol. 135, p. 109864, 2020.
[13] S. Dutta and S. K. Bandyopadhyay, "Machine learning approach for confirmation of covid-19 cases: Positive, negative, death and release," MedRxiv, 2020.
[14] A. Tomar and N. Gupta, "Prediction for the spread of COVID-19 in India and effectiveness of preventive measures," Science of The Total Environment, vol. 728, p. 138762, 2020.
[15] C. Xu, "A Comparative Study: Time-Series Analysis Methods for Predicting COVID-19 Case Trend," ed, 2021.
[16] H. Baalbaki et al., "Fighting against COVID-19: Who Failed and Who Succeeded?,", Journal of Computer and Communications, vol. 10, no. 4, pp. 32-50, 2022.
[17] A. B. Said, A. Erradi, H. A. Aly, and A. Mohamed, "Predicting COVID-19 cases using bidirectional LSTM on multivariate time series," Environmental Science and Pollution Research, vol. 28, no. 40, pp. 56043-56052, 2021.
[18] J. MacQueen, "Some methods for classification and analysis of multivariate observations," 1967.
[19] D. Arthur and S. Vassilvitskii, (10 June 2022). *k-means++: The Advantages of Careful Seeding*, Stanford InfoLab, Technical Report 2006. [Online]. Available: http://ilpubs.stanford.edu:8090/778/
[20] R. C. Staudemeyer and E. R. Morris, "Understanding LSTM--a tutorial into long short-term memory recurrent neural networks," arXiv preprint arXiv:1909.09586, 2019.