Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
5th International Conference on Computer Science and Computational Intelligence 2020

Predict Mortality in Patients Infected with COVID-19 Virus Based on Observed Characteristics of the Patient using Logistic Regression

Bernhard O. Josephusa, Ardianto H. Nawirb, Evelyn Wijayae, Jurike V. Moniagad*, Margaretha Ohyverc

a,b,c,d Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480
b Statistics Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480

d Keywords: Covid-19, Logistic Regression, Mortality

Abstract

The spread of COVID-19 has made the world a mess. Up to this day, 5,235,452 cases confirmed worldwide with 338,612 death. One of the methods to predict mortality risk is machine learning algorithm using medical features, which means it takes time. Therefore, in this study, Logistic Regression is modeled by training 114 data and used to create a prediction over the patient’s mortality using nonmedical features. The model can help hospitals and doctors to prioritize who has a high probability of death and triage patients especially when the hospital is overrun by patients. The model can accurately predict with more than 90% accuracy achieved. Further analysis found that age is the most important predictor in the patient’s mortality rate. Using this model, the death rate caused by COVID-19 could be reduced.

© 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 5th International Conference on Computer Science and Computational Intelligence 2020

1. Introduction

The new coronavirus family, COVID-19, is a zoonosis-type virus that spreads from animals to humans. COVID-19 first appeared in Wuhan City, China, at the end of 2019 and quickly spreads to various countries, including Indonesia1,2. Up until the day this paper writing, 5,235,452 people all around the world have been infected, 2,072,768
people have been recovered, and 338,612 people had died. In Indonesia’s case, 21,745 people have been infected, 5,249 people have been recovered, and 1,351 people had died\(^3\).

Looking at how quickly the virus spread, scientist around the world trying to collect clinical features from many infected patients by using a machine learning-based prognostic model with clinical data in Wuhan and finally conclude that these three key clinical features, \textit{lactic dehydrogenase} (LDH), \textit{lymphocyte} and \textit{High-sensitivity C-reactive protein} (hs-CRP) play a huge role in a severe COVID-19 patients survival\(^4\). However, those clinical features need a doctor to get the appropriate result. Only doctor can examine patient and determine the severity using the three features.

Artificial Intelligence (AI) has shown being an effective tool in predicting medical conditions\(^5\). In this study, a predictive algorithm based on Artificial Intelligence (AI) was used to predict the death of COVID-19 patients. The algorithm used observed characteristics of the patients to predicts the mortality risk. With the help of the algorithm, the hospital can triage appropriate patients especially when the hospital is overcrowding.

The proposed algorithm used in this paper is Logistic Regression. Logistic regression is a mathematical model which describe the relation between one or more independent variables and a qualitative dependent variable. This dependent variable has two or more categories. If the dependent variable has two categories, then the model is called a binary logistic regression model. If the dependent variable has more than two categories, the model is called a multinomial or ordinal logistic regression model\(^6,7\). Other modelling approaches are possible also, but the most popular of these approaches is the logistic model, which is estimated by maximum likelihood\(^7\). What makes it so popular is the logistic function, which describes the mathematical form on which is an extremely flexible and easily used function\(^8\).

Logistic Regression has been used in several studies related to COVID-19. This method has been used to predict the total number of people with COVID-19\(^9\) to model the spread of COVID-19 in China\(^10\), and to predict the trend of the COVID-19 epidemic\(^11\). XGBoost machine learning algorithm also was used to predict the mortality risk and used the LDH, \textit{lymphocyte}, and hs-CRP\(^4\). With these three features, patients have to do medical test. Therefore, this paper used Logistic Regression to predict mortality risk with several nonmedical features.

The goal of this paper is to create a binary logistic regression model that can accurately predict mortality risk in COVID-19 patients to help hospitals and medical facilities prioritize the patients who have the highest death probability and give the appropriate treatment. We hope this study could minimize mortality due to COVID-19 and led more people to aware of the disease especially for the people who have a high probability of death classified by our model.

2. Methods

2.1. Logistic Regression

Assume there is one independent variable, \(X\), and one dependent variable, \(Y\), that have two categories. Let \(P(Y = 1|X = x) = 1 - P(Y = 0|X = x)\). The logistic regression model is

\[
\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}
\]

(1)

The extended model in (2) applies for multiple binary logistic regression.

\[
\pi(x_i) = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}}}
\]

(2)

The \(\beta_0, \beta_1, \ldots, \beta_k\) are the parameters for the model. The estimation for the parameters determined by the maximum likelihood estimation. The first step to estimate the parameter is defining the likelihood function. Assume there are \(Y_1, Y_2, \ldots, Y_N\) binomial random variables. As the observations are assumed to be independent, the likelihood function for these binomial random variables can be seen in the following formula.

\[
L(\beta) = \prod_{i=1}^{N} \frac{n_i!}{y_i!(n_i-y_i)!} \pi(x_i)^{y_i}(1-\pi(x_i))^{n_i-y_i}
\]

\[
L(\beta) = \prod_{i=1}^{N} \frac{n_i!}{y_i!(n_i-y_i)!} \pi(x_i)^{y_i}(1-\pi(x_i))^{n_i-y_i}
\]

(3)

Taking the natural logarithm of \(L(\beta)\),

\[
\ln(L(\beta)) = \sum_{i=1}^{N} y_i \ln[\pi(x_i)] - (1 - y_i) \ln[1 - \pi(x_i)]
\]

(4)

Differentiating \(L(\beta)\) with respect to \(\beta_0, \beta_1, \ldots, \beta_k\). Set the result of this differentiation equal to zero. This result is not a closed form formula, so we require iterative methods to get the estimated coefficients, for example the Iterative Weighted Least Squares method.
2.2. A Chi-Square Test for Independence

A Chi-Square test for independence is a test used to check the independent relation between two categorical variables. This test will make use of contingency tables, i.e. tables with cells corresponding to cross-classifications of attributes or events. The hypothesis is

\[ H_0: \text{The two categorical variables are independent of each other.} \]
\[ H_1: \text{The two categorical variables are not independent.} \]

The chi-square test statistic for this test is

\[ \chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \]  

The count of the elements in cell (i,j), that is, the cell in row I and column j (where \( i = 1, 2, \ldots, r \) dan \( j = 1, 2, \ldots, c \)), is denoted by \( O_{ij} \). The \( E_{ij} \) is the expected count in cell (i, j) which defined by

\[ E_{ij} = \frac{R_i C_j}{n} \]

The \( R_i \) and \( C_j \) are the total count for row i and the total count for column j.

2.3. Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA) is a statistical method for determining the existence of differences among several population means. In one-way ANOVA, to analyze variation towards the goal of determining possible differences among the group means, you partition the total variation into variation that is due to differences among the groups and variation that is due to differences within the groups. The within-group variation \( (SS_W) \) measures random variation. The among-group variation \( (SS_A) \) measures differences from group to group. The symbol \( n \) represents the number of values in all groups and the symbol \( c \) represents the number of groups.

The hypothesis for doing ANOVA analysis is

\[ H_0: \mu_1 = \mu_2 = \ldots = \mu_r \]
\[ H_1: \text{Not all } \mu_i (i = 1,2,\ldots,r) \text{ are equal.} \]

2.4. Dataset

In this paper, the dataset is obtained from Kaggle with a total of 1085 cases/data and 25 features including gender, age, and location from January 20\textsuperscript{th} to February 25\textsuperscript{th}, 2020. Figure 1 shows how is the data pre-processing done. First, unnamed and useless columns such as id, summary, source, and link are removed. Data with no target value are removed from the dataset leaving 222 data. To deal with missing values, a column that has more than 60\% missing values is removed, a row with missing categorical value is removed, and data imputation is performed.

Next, the date of the symptom onset and date of hospital visits are combined into the time gap between symptom onset and hospitalization. The combined value of hospitalization time gap with negative value - which means hospital visit date preceded symptom onset date - is removed. The result of this process leaving 114 valid data.

3. Result and Discussion

Feature selection is useful for building simpler and more comprehensible models, improving data-mining performance, and preparing clean, understandable data\textsuperscript{12}. Some of the methods include the ANOVA test and the chi-square test. The features of the dataset contain both numerical and categorical data, therefore both methods are used. Both methods will give a corresponding p-value for each feature which will be used to compare to a significance value.

The initial features are Country, Gender, Age, Hospitalization Time Gap, From Wuhan, and Visit Wuhan. First, Chi-square is used to calculate the relation between a categorical feature and categorical target\textsuperscript{13}, in this case, Country, Gender, From Wuhan, and Visit Wuhan.
Fig. 1. Data Preprocessing Flow

Table I shows the p-values of each categorical feature. Let define significance value $\alpha = 0.05$. A feature that has p-values lower than $\alpha$ is taken, which is Country, Visit Wuhan, and From Wuhan.

| Country | Visit Wuhan | From Wuhan | Gender |
|---------|-------------|------------|--------|
| p-values | $1.6 \times 10^{-24}$ | $2.4 \times 10^{-5}$ | $3.0 \times 10^{-5}$ | $3.9 \times 10^{-5}$ |

Next, ANOVA is used to calculate the relation between a numerical feature and categorical target, which is Age and Hospitalization Gap. Table II shows the p-values of each numerical feature.

| Age | Hospitalization Time Gap |
|-----|---------------------------|
| p-values | $6.74 \times 10^{-15}$ | $2.85 \times 10^{-5}$ |

Using the same significance value and hypothesis as in Chi-square test, all numerical feature is taken. Taking all the selected features, the final features are Age, Hospitalization Gap, Country, Visit Wuhan, and From Wuhan.

The final dataset is trained using Logistic Regression using the LogisticRegression package available in python (sklearn). The dataset is randomly split into data train and data test with ratio 70:30, where 70% of the final dataset is used for data train and 30% of the dataset is used for data tests. As the dataset is small, a Liblinear solver is used for the training. The model is then evaluated using Precision, Recall, F1 Score, Confusion Matrix, and Area Under the Receiver Operating Characteristics (ROC) Curve.

Out of 1085 cases/data, 114 data are picked with Age, Hospitalization Gap, Location, and Country as the predictor. The final data composed of 80 patients recovered and 34 patients died. The training dataset is randomly picked containing 70% of the data and 30% of the data is used for testing. The performance evaluation after testing is shown and discussed below.
Figure 2 presents the confusion matrix for the data. From this figure, we know that the model can 100% accurately predict the patient’s death and 96% accurately predict a patient’s survival, as only 1 case is missed. Precision and recall are also used to evaluate performance. Table III shows the precision, recall, f1 score, and corresponding support for each class. The precision score of death is the only one that is lower than 0.9, while survival, accuracy, macro averages, and weighted averages score are larger than 0.9. This indicates the good performance of the model.

|            | Precision | Recall | F1 Score | Support |
|------------|-----------|--------|----------|---------|
| Death      | 0.86      | 1.0    | 0.92     | 6       |
| Survival   | 1.0       | 0.97   | 0.98     | 29      |
| Accuracy   | -         | -      | 0.97     | 35      |
| Macro Avg  | 0.93      | 0.98   | 0.95     | 35      |
| Weighted Av| 0.98      | 0.97   | 0.97     | 35      |

While the result above takes only one threshold for the evaluation, ROC Curve plots the True Positive Rate, as y-axis, and False Positive Rate, as x-axis, but with various thresholds. To construct a ROC curve, we calculate the True Positive Rate (TPR) and False Positive Rate (FPR) for each threshold with the following formula. All the variables (True Positive, False Negative, False Positive, True Negative) used in the formula can be taken by constructing a confusion matrix, similarly as in Figure 2. By using the `roc_curve` package in Python, the TPR, FPR, and threshold are calculated automatically. Table IV below is the result.

|            | FPR   | TPR   |
|------------|-------|-------|
| FPR        | 0     | 0     |
| TPR        | 0     | 0.034 |
| Threshold  | 1.995 | 0.995 |

The calculated TPR and FPR values are then plotted to the graph to make a ROC Curve. Figure 3 shows the ROC curve graph.
The point here is to calculate the area under the curve score. It ranges from 0.0 to 1.0. Higher the score, better the model. Surprisingly, the area under the curve score of the trained model is 1.0 which means the model has a very good prediction of COVID-19 patients mortality risk.

Feature importance is used to find out which features are the most important in predicting the outcome/output. Figure 4 shows the result of the calculated feature importance score of each feature, from the highest score to the lowest. Age seems to be the most important feature in predicting the patient’s survival, following by hospitalization time gap, from Wuhan, country, and visit Wuhan.

4. Conclusion

COVID-19 patient’s mortality risk could be predicted by developing a Logistic Regression model with Age, Hospitalization Time Gap, From Wuhan, Country, and Visit Wuhan as the predictors. The model developed has shown a good performance based on all metrics. It can help hospitals prioritize patients who really in need and reduce the mortality rate. However, the predictive model just being feed by a small amount of data, which may result in a lack of recognizing the pattern. In future studies, gathering more data for training is expected.

References

1. Organization WH. Laboratory Testing for Coronavirus Disease 2019 (COVID-19) in Suspected Human Cases: Interim Gui. World Health Organization; 2020.
2. Xu XW, Wu XX, Jiang XG, Xu KJ, Ying LJ, Ma CL, et al. Clinical Findings in a Group of Patients Infected with the 2019 Novel Coronavirus (SARS-Cov-2) Outside of Wuhan, China: Retrospective Case Series. BMJ. 2020 February.

3. Google News. [Online]; 2020 [cited 2020 March 31. Available from: https://www.google.com/covid19-map/.

4. Yan L, Zhang HT, Xiao Y, Wang W, Guo Y, Sun C, et al. Prediction of Survival for Severe Covid-19 Patients with Three Clinical Features: Development of a Machine Learning-Based Prognostic Model with Clinical Data in Wuhan. medRxiv. 2020.

5. Pourhomayoun, Mohammad, Shakibi M. Predicting Mortality Risk in Patients with COVID-19 Using Artificial Intelligence to Help Medical Decision-Making. medRxiv. 2020.

6. Ohyver M, Moniaga JV, Yunidwi , Restisa K, Setiawan MII. Logistic Regression and Growth Charts to Determine Children Nutritional and Stunting Status: A Review. In Procedia Computer Science; 2017; Denpasar. p. 232-241.

7. Allison PD. Logistic regression using SAS: Theory and application. 2nd ed.: SAS Institute; 2012.

8. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression. 3rd ed.: John Wiley & Sons; 2013.

9. Batista M. Estimation of the final size of the coronavirus epidemic by the logistic. medRxiv. 2020 February.

10. Shen CY. Logistic Growth Modelling of COVID-19 Proliferation in China and Its International Implications. International Journal of Infectious Diseases. International Journal of Infectious Diseases. 2020 July; 96.

11. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, et al. Feature Selection: A Data Perspective. ACM Comput Surv. 2017 December.

12. Sharpe, Donald. Chi-Square Test is Statistically Significant: Now What? Practical Assessment, Research, and Evaluation. 2015; 20.