

Review Article

Etiology management and COVID-19 death forecast using artificial intelligent based Bayesian learning approach

Mohd Saqib¹, Saleem M. Khan^{2*}, Danish Suhail³

¹Department of Mathematic and Computing, Indian Institute of Technology (Indian School of Mines), Dhanbad, Jharkhand, India

²Department of Community Medicine, JN Medical College, AMU, Aligarh, U.P. India

³Surveillance Medical Officer, NPSP- World Health Organisation, India

Received: 05 November 2020

Revised: 18 January 2021

Accepted: 01 February 2021

*Correspondence:

Dr. Saleem M. Khan,

E-mail: saleemmohdkhan@gmail.com

Copyright: © the author(s), publisher and licensee Medip Academy. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT

In 2020, coronavirus disease 2019 (COVID-19), caused by the SARS-CoV-2 (severe acute respiratory syndrome corona virus 2) coronavirus, has become a worldwide natural disaster and has taken form of pandemic. Due to this unforeseen pandemic, humanity had faced many challenges and realized the capability of a human being is not enough to survive in such circumstances. Here, artificial intelligence (AI) comes into the picture which can perform beyond human abilities and enhance the capabilities of a human especially the doctors and health care system. The proposed work is a survey of the areas of COVID-19 where AI can play a big role. There are some areas: death cases prediction, virus progression prediction, disease detection, and drug discovery. In this study we have discussed above mentioned issues and presented a Bayesian learning model to forecast death cases in India. The proposed model predicted with good accuracy and also takes care of uncertainty of predictions. We have fitted Bayesian learning on n-polynomial. This is a completely mathematical model in which we have successfully incorporated with prior knowledge and posterior distribution enables us to incorporate more upcoming data without storing previous data. Our forecast in this study is based on the public datasets provided by John Hopkins University. We are concluding with further evolution and scope of the proposed model.

Keywords: COVID-19 pandemic, Bayesian leaning, Prediction, Etiology management

INTRODUCTION

In late December 2019, a group of patients were come up with an unknown etiology to the hospitals having symptoms of pneumonia. Later on the first case of novel coronavirus was reported in the city of Wuhan in Hubei province in central China.¹ After taking basic understanding of virus, medical experts given a name as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and the name of disease caused by this virus is coronavirus disease 2019 (COVID-19).² The cases of COVID-19 pandemic are rapidly growing. Till the date 30

April 2020, worldwide we have 3,251,587 confirmed and 229,832 death cases due to this hazardous pandemic, COVID-19.

In India, the first laboratory-confirmed case of COVID-19 was reported from Kerala on 30 January 2020 and as of 30 April 2020, a total of 33,931 cases and 943 deaths were reported in India.³ To tackle this ongoing pandemic and such events in the future where the lives of millions of people are at high risk, we need a strong health care system and technology is the means of making a way to a panacea. Despite several research works and their documentation, there is a huge opportunity for utilization of AI, machine

learning, and data science in this field, due to novelty of the root cause. That is why some prediction methods are required to estimate the magnitude of death cases and masses of studies on distinct methods of forecasting are represented so far.⁴ In some papers, authors estimate the possible number of infected cases in India using long short-term memory (LSTM).^{4,5} Same as in⁶, study represented virus progression and forecast using the same algorithm for Canada and compare with United State (US) and Italy. A different research also take place on the cases of India but using different methods, an autoregression integrated moving average model (ARIMA) and Richard's model.⁷ Moreover predictions, some mathematical models are also estimated the effects of lockdown and social-distancing in India in practical scenario.⁸ The studies represented so far are based on inadequate data at initial stage without any measurement of uncertainty. These models are developed with good accuracy but as well as data become available, those entire algorithm will not able to survive without little bit of evaluations due to dynamic nature of pandemic escalation of the COVID-19. So, a distribution based learning model will be more promising rather than doing point estimation. Bayesian Learning is very well known method of making any prediction based on our prior knowledge.⁹ In the proposed study we are formulating Bayesian learning regression with polynomial of n-degree.

Etiology of COVID-19 and artificial intelligence

AI is the bigger combination of many mathematical, statistical, and computer science algorithms. Followings are the main parts of the AI.

Machine learning

Machine learning is the base of any AI technique. It has various types of algorithms which can be classified into three types: supervised, unsupervised, reinforcement learning which can solve three types of problems (Figure 1); regression, classification, and clustering. In the case of COVID-19, we are facing various types of problems which may categories in above mentioned three categories of machine learning problems. For example, regression models like linear regression, polynomial regression, support vector regression, etc. can be used to predict the virus progression rate. Classification models like decision tree, support vector machine, random forest, etc. can be used to classify the data in two main categories of patients-normal, and infected by corona. Clustering models like K-nearest neighbour (KNN) can be utilized in making the category of the unlabelled data e.g. categories the fake and real article authenticating from the World Health Organization (WHO) spreading by social media.

Deep learning

Deep learning is the extension of machine learning. Various complex algorithms fell into deep learning like artificial neural network (ANN), convolutional neural network (CNN), etc. which resolve all the kind of problems

handle by machine learning but it can enable us to work with big data, images, audio, video, and other raw multimedia data instead of working with only well-organized structured tabular and textual data. Nowadays, during this pandemic, medical organizations are generating various lab reports, X-ray and computed tomography (CT) scan very fast. All these can be given new insight regarding COVID-19 using deep learning.^{10,11}

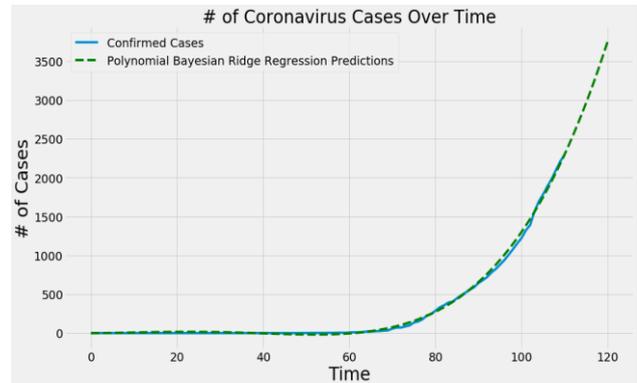


Figure 1: Actual and forecast values of death cases in India.

Data mining

Data mining is a broader part of knowledge discovery of database (KDD) and used extract rules, the association among entities, pattern from the vast data. During this pandemic, various newspapers, social media, blogging websites, and other various online platforms actively spreading news and articles regarding the virus, hygiene guidelines, progression, etc. which must be authenticated from reliable resources such as WHO. But in the era of social media detecting, fake news and articles become so hard. Here Data Mining plays an important role and finds authenticate article which is associated with WHO.¹²

Natural language processing

Natural language processing (NLP) gives understandability of the human languages which may be in text form or audio on the computer. For example a research work presents an algorithm for automated diagnosis of pertussis using audio signals by analyzing cough and whoop sounds.¹² In our case, sounds of cough may use to determine if the patient is COVID positive or negative.

Computer vision

AI makes a computer to mimic human then it's become necessary that how a computer sees and identify the objects like a human. Computer Vision provides this facility.¹³ During this pandemic when most of the cities are struggling to maintain lockdown and people are violating instruction of social distancing, Computer Vision provides access to monitor remotely via CCTV cameras and

identifying them. **Error! Reference source not found.** demonstrating the scenario.

COVID-19 death prediction

Datasets

The datasets collected from Johns Hopkins University are used in the studies.¹⁴ The datasets accessed on 11 September 2020. It provides several fatalities and registered patients by the end of each day. The dataset is available in the time series format with date, month, and year so that the temporal components are not neglected. A wavelet transformation is applied to preserve the time-frequency components and it also mitigates the random noise in the dataset.¹⁵ The datasets are divided into two parts training (80%) and testing (20%) datasets.

Model formulation

One of the very basic approaches to make a prediction is another version of linear regression is Polynomial regression in which the relationship between independent and dependent variables is an n-degree polynomial. Mathematically it can represent as follows:

$$\begin{aligned}
 f(X) & & 1 \\
 &= \beta_0 + \beta_1 x_1^1 \\
 &+ \beta_2 x_2^2 + \dots \\
 &+ \beta_n x_n^n + \epsilon
 \end{aligned}$$

Or,

$$\begin{aligned}
 f(X) & & 2 \\
 &= \beta_0 + \sum_{i=1}^n \beta_i x_i^i \\
 &+ \epsilon
 \end{aligned}$$

Where β_i is the coefficient and ϵ the measurement error which is

$$\epsilon \sim N(0, \sigma^2) \quad 3$$

$f(X)$ is our polynomial model and to develop a good model we need to tuning, β_i So that following loss function with L^2 Regularization (Ridge Regularization) will be as minimum as possible

$$\begin{aligned}
 \beta &= L(y_i, x_i) & 4 \\
 &= \sum_{i=1}^n (y_i - f(x_i))^2 \\
 &+ \lambda \sum_{i=1}^n \beta_i^2
 \end{aligned}$$

Where, the first part of the equation 4 is the residual sum of squares (RSS), the difference between actual value (y_i) and predicted value ($f(x_i)$) of the i^{th} observation. λ is the regularization term, deciding how much regularize the β_i .

Now, for the best fitting our aim to minimize the β by tuning coefficients, β_i .¹⁶ According to the maximum likelihood estimate of β which reduces the $L(y_i, x_i)$ is

$$\hat{\beta} = (x^T x)^{-1} x^T y \quad 5$$

Now, instead of a vector of coefficients, we have a single value $\hat{\beta}$, in \mathbb{R}^{p+1} .¹⁷ Here Bayesian Regression (BR) comes into the picture. In the BR, instead of predicting value mentioned as above, it used probabilistic distribution to estimate the value of y_i and its follow the following syntax

$$y_i \sim N(\beta^T X, \sigma^2) \quad 6$$

So,

$$\begin{aligned}
 p(y | X, \beta, \sigma^2) & & 7 \\
 \propto \frac{1}{n^{1/2} \sqrt{\sigma^2}} e^{-\frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta)}
 \end{aligned}$$

From conjugate prior distribution,

$$\begin{aligned}
 (y - X\beta)^T (y - X\beta) & & 8 \\
 &= (y - X\hat{\beta})^T (y - X\hat{\beta}) \\
 &- (\beta - \hat{\beta})^T (X^T X) (\beta - \hat{\beta})
 \end{aligned}$$

The equation 7 is re-written as

$$p(y | X, \beta, \sigma^2) \propto \frac{1}{\frac{n}{2}\sqrt{\sigma^2}} \exp\left(-\frac{vs^2}{2\sigma^2}\right) (\sigma^2)^{-\frac{n-v}{2}} \exp\left(-\frac{1}{2\sigma^2}(\beta - \hat{\beta})^T(X^T X)(\beta - \hat{\beta})\right) \tag{9}$$

coefficients in vector β . This suggests a form for the prior distribution is

$$p(\beta, \sigma^2) = p(\sigma^2)p(\beta | \sigma^2) \tag{10}$$

Where, $vs^2 = (y - X\hat{\beta})^T(y - X\hat{\beta})$ and $v = n - k$, n is the number of observations, and k is the number of

After the formulation of the prior distribution, now we need to generate posterior distribution, which can be formulated as follow (from equations 7, 9 and 10)

$$p(\beta, \sigma^2 | y, X) \propto p(y | X, \beta, \sigma^2) p(\beta | \sigma^2) p(\sigma^2) \tag{11}$$

$$p(\beta, \sigma^2 | y, X) \propto \frac{1}{\frac{n}{2}\sqrt{\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)} (\sigma^2)^{-\frac{n-v}{2}} e^{-\frac{1}{2\sigma^2}(\beta-\mu_0)^T\Lambda_0(\beta-\mu_0)} (\sigma^2)^{-(a_0-1)} e^{\left(\frac{b_0}{\sigma^2}\right)}$$

Where Λ_0 is ridge regression used to overcome the problem of multicollinearity normally occurring when the model has large numbers of parameter.¹⁸ And it is equal to

$$\Lambda_0 = cI \tag{12}$$

Now we are ready to estimate the probability of y on given conditions (m) using Bayes Theorem as

$$p(y | m) = \frac{p(\beta, \sigma | m)p(y | X, \beta, \sigma, m)}{p(\beta, \sigma | y, X, m)} \tag{14}$$

In equation 12, $c = \sum_{i=1}^n \beta_i$ and I is an identity matrix of $n \times n$. Now, the posterior mean (μ_n) can be represented in the term of $\hat{\beta}$ and prior mean μ_0 and for the Bayesian learning other can be upgraded as follows

$$\mu_n = (X^T X + \Lambda_0)^{-1}(X^T X \hat{\beta} + \Lambda_0 \mu_0) \tag{13}$$

Where m is the marginal likelihood and prior density, here, m is $p(y | X, \beta, \sigma)$.

Implementation

We have implemented the model in Python.3 using scikit-learn and Spyder to write the code. We have implemented the proposed model on India’s death cases and found root mean square error (RMSE) is 48.11. Figure 1 represents the actual and forecast magnitude of death cases in India

RESULTS AND DISCUSSION

It is demanding to construct a model to predict the death cases of COVID-19 situations. So many researchers are struggling to find and implementing such models with optimal parameters and unknown variables which lead them to uncertainty. Bayesian learning model is different from all the studies published or at least discussed in the literature survey because of its nature of making estimation. It is a complex mathematical model that more focused to discover distribution instead of making a single value linear prediction of the dependent variable and this feature makes it more promising. In the Table 1, we have given criteria used in evaluation of the model.

Table 1: Model evaluation criteria.

Criteria	Value
MAE	41.90
MSE	2314.88
RMSE	48.11
NRMSE	0.02
MEAN	1298.99
SD	535.67
NSD	0.41
R2_test (%)	99.22
R2_train (%)	98.53

CONCLUSION

After all the above studies, we have seen AI can play dominating role during this pandemic. Not only help us to predicting progression and positive patients of Covid 19 but also keep high the level of different management and monitoring system like lockdown, social distancing, etc. It also gives new insight of the pandemic data using ANN, CNN, and Bayesian learning approach.

Funding: No funding sources

Conflict of interest: None declared

Ethical approval: Not required

REFERENCES

- World Health Organization. Novel coronavirus - China. Available at: <http://www.who.int/csr/zxcvXDdon/12-January-2020-novel-coronavirus-china/en/>. Accessed on: 21 April 2020.
- World Health Organization. Available at: [https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-\(covid-2019\)-and-the-virus-that-causes-it](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it). Accessed on: 21 April 2020.
- Ministry of Health & Family Welfare, Government of India. COVID-19 India Update. Available at: <https://www.mohfw.gov.in/>. Accessed on: 21 April 2020.
- Tomar A, Gupta N. Science of the Total Environment Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. *Sci Total Environ.* 2020;728:138762.
- Ranjan R. Predictions for COVID-19 outbreak in India using Epidemiological models Predictions for COVID-19 outbreak in India. 2020.
- Kumar V, Chimmula R, Zhang L. Time Series Forecasting of COVID-19 transmission in Canada Using LSTM Networks. *Chaos, Solitons Fractals Interdiscip J Nonlinear Sci Nonequilibrium Complex Phenom.* 2020:109864.
- Nanda C, Space H, Centre A. Forecasting COVID-19 impact in India using pandemic waves Nonlinear Growth Models. 2020.
- Arti MK. Modeling and Predictions for COVID 19 Spread in India. 2020.
- Bayesian linear regression. Wikipedia. Available at: https://en.wikipedia.org/wiki/Bayesian_linear_regression. Accessed on: 21 April 2020.
- Kumar P, Kumari S. Detection of coronavirus Disease (COVID-19) based on Deep Features. Preprints. 2020;2020030300.
- Lomoro P, Verde F, Zerboni F. COVID-19 pneumonia manifestations at the admission on chest ultrasound, radiographs, and CT : single-center study and comprehensive radiologic literature review. *Eur J Radiol Open.* 2020;7:100231.
- Asr FT, Taboada M. Big Data and quality data for fake news and misinformation detection. *Big Data Soc.* 2019;1-14.
- Arbor A, Lynch JP. A Computer Vision Framework for Human User Sensing in Public Open Spaces A Computer Vision Framework for Sensing Human Users in Public Open Spaces. DFHS'19, New York, NY, USA. 2020;27-30.
- COVID-19 Datasets from Johns Hopkins University. Available at: <https://github.com/CSSEGISandData/COVID-19>. Accessed on: 11 May 2020.
- Xu Y, Weaver JB, Healy MD. Wavelet transform domain filters: a spatially selective noise filtration technique. *IEEE Trans image Process.* 1994;3(6):747-58.
- Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer. 2009.
- Bayesian Linear Regression Models with PyMC3. Quantstart. Available at: <https://www.quantstart.com/articles/Bayesian-Linear-Regression-Models-with-PyMC3/>. Accessed on: 11 May 2020.
- Gruber M. *Improving Efficiency by Shrinkage: The James–Stein and Ridge Regression Estimators.* Boca Rat CRC Press. 1998;7-15.

Cite this article as: Saqib M, Khan SM, Suhail D. Etiology management and COVID-19 death forecast using artificial intelligent based Bayesian learning approach. *Int J Community Med Public Health* 2021;8:1542-6.