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.
Forecasting and classification of new cases of COVID-19 before vaccination using decision trees and Gaussian mixture model

Monia Hamdia, Inès Hilali-Jaghdamb, Bushra Elamin Elnaimc, Azhari A. Elhagd

a Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
b Department of Computer Sciences and Information Technology, Applied College, Princess Nourah bint Abdulrahman University, P. O. Box 84428, Riyadh 11671, Saudi Arabia
c Department of Computer Science College of Science and Humanities in Al-Sulail, Prince Sattam bin Abdulaziz University, Saudi Arabia
d Department of Mathematics, College of Science, Taif University, P.O. Box 11099, Taif, 21944, Saudi Arabia

Received 5 April 2022; revised 16 June 2022; accepted 3 July 2022
Available online 08 July 2022

KEYWORDS
Gaussian Mixture Model (GMM); Decision Tree (DT); Machine Learning (ML); Chi-Squared Automatic Interaction Detection (CHAID)

Abstract Regarding the pandemic taking place in the world from the spread of the Coronavirus pandemic and viral mutations, the need has arisen to analyze the epidemic data in terms of numbers of infected and deaths, different geographical regions, and the dynamics of the spread of the virus. In China, the total number of reported infections is 224,659 on June 11, 2022. In this paper, the Gaussian Mixture Model and the decision tree method were used to classify and predict new cases of the virus. Although we focus mainly on the Chinese case, the model is general and adapted to any context without loss of validity of the qualitative results. The Chi-Squared ($\chi^2$) Automatic Interaction Detection (CHAID) was applied in creating the decision tree structure, the data has been classified into five classes, according to the BIC criterion. The best mixture model is the E (Equal variance) with five components. The considered data sets of the world health organization (WHO) were used from January 5, 2020, to 12, November 2021. We provide numerical results based on the Chinese case.

© 2022 THE AUTHORS. Published by Elsevier BV on behalf of Faculty of Engineering, Alexandria University
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

The analysis of the data that COVID-19 is increasing day by day since it appeared in China on December 30, 2019. This is due to the virus’s seriousness, its damages in losses in human lives, and the global economy incurred. This appeared in the research of many researchers, where [1] compared the Artificial
neural networks (ANN) and the logistic regression (LR) models using the data of covid-19 and it appeared that the performance of ANN is better than LR, where the classification accuracy rate (CAR) was higher at 85.6% comparing with 80.8%. This article [2] discusses the short-term prediction of the data novel coronavirus (COVID-19) in India for those who infected, recovered, and active cases using a machine-learned hybrid Gaussian (MLHG) and time series method. This paper [3] attempted to predict the short-term dynamics of data of patients hospitalized with COVID-19 in Italy. Both single time series forecasts and feasible hybrid methods and combinations were used. The results showed that the best single models were Neural Network Auto-Regressive (NNAR) and Auto-Regressive Integrated Moving Average (ARIMA), and the most accurate were the hybrid models NNAR-TBATS* and ARIMA-NNAR (see Fig. 1).

The authors of [4] studied different statistical models represented in the time series models, which are easily available from the R package. These models were applied to the data of the Coronavirus, as the number of diseased cases, i.e., the confirmed cases and the number of deaths, and the number of cases that recovered at the 90% confidence level for some time were expected. Forecasting to estimate the uncertainty about point predictions. These different models have been applied to the data presented for the data of COVID-19 pandemic in the United State of America and Italy. In this paper, it was noted that all the models that were used projected the future numbers of the confirmed cases, deaths, and recoveries for COVID-19 data with high accuracy. Both the ARIMA time series model and the cube smoothing slice models have lesser prediction errors and smaller prediction periods, compared to the Holt model and the trigonometric exponential smoothing model (TESM) using Box-Cox transform models. Thus, the preceding two models were better than the last two. In this article [5] time series models were used to predict COVID-19 infection with data from Saudi Arabia, and the prediction results obtained showed that the ARIMA model has had a less significant projecting error in most evaluations. Here also noted that the time series models are accessible for use at any time and in any country, with reasonably accurate forecasts. The cubic spline, Holt, and Holt-Winters models also presented well in many of the experiments conducted, and the results of this study can be used when studying further models, for instance deep learning models. In the current article [6] an estimate of the increase in all-cause mortality in Brazil and Spain is presented using monthly mortality data from 2015 to 2019.

Besides the reported COVID-19 deaths, the additional deaths also serve as a tool to assess the impact of the ongoing epidemic. Projects had been made for the total deaths that occurred in 2020 and the first third of 2021 in the lack of the COVID-19 pandemic. To achieve this goal, the accuracy of the three distinct prediction methods extensively utilized in data enumeration was studied. It has been observed that the ARIMA model class gives the best results compared to Exponential Smoothing (ETS) and generalized linear models (GLM). Under the GLM category, autoregressive Poisson models and negative binomial models, as well as logarithmic shift on parameters were considered to what extent the best performance of the ARIMA and ETS models were observed. Those that belong to the GLM category. Mortality data are known to exhibit excessive scatter, and this may be due to the increased performance of other models, at least in the countries reviewed in this paper. Unquestionably, a comparative assessment of these models for a like study of previous epidemics in other countries is an interesting task that needs to be solved in the future. This paper aimed to use the Gaussian Mixture Model and the decision tree method to model the new cases in China as machine learning methods to classify and predict new cases of coronavirus in China.

* TBATS This is an abbreviation of the main characteristics of the model: T: Trigonometric seasonality B: Box-Cox

![Flowchart](image.png)

Fig. 1 Repeat E-step and M-step until convergence [21].
transformation A: ARIMA errors T: Trend S: Seasonal components (see Fig. 2).

2. Decision tree (DT)

Decision trees are used to build automated predictive models for machine learning, data mining, and statistics [7,8]. This method is called “learning through a decision tree” and allows you to predict the value of an object, considering the collected observations about it [9]. DT schematically shows the possible consequences of a series of related decisions [10]. This approach allows individuals and organizations to weigh the different options for action, considering the costs involved, the level of likelihood, and the perceived “value” of each option (that is, the average over each outcome) [11]. Today there are many learning algorithms: ID3, CART, C4.5, C5.0, CHAID, CN2, and others. The most popular are:

ID3 (Iterative Dichotomize 3). The algorithm allows you to work only with a discrete target variable. Decision trees built based on ID3 are qualified. The number of descendants in a node is unlimited. The algorithm does not work with missing data.

C4.5. “Advanced” version of ID3, supplemented by the ability to work with missing attribute values. In 2008, Spring Science conducted a study and found that C4.5 is the most popular Data Mining algorithm.

CART (Classification and Regression Tree) [12]. The algorithm solves the problems of classification and regression, as it allows the use of discrete and continuous target variables. CART builds trees with only two children per node.

3. Gaussian mixture model (GMM)

GMM is defined as a type of machine learning algorithm [13,14]. GMM is used to predict and classify data into different classes based on the probability distribution [15]. GMM is used in different areas, such as marketing, finance, and so on [16]. In the Gaussian mixture model, the distribution can be represented as.

\[
p(x) = \sum_{j=1}^{k} w_j N(x | \mu_j, \Sigma_j)
\]

\[
\sum_{j=1}^{k} w_j = 1, 0 \leq w_j \leq 1
\]

Where is the prior probability of the \(j\)th gaussian.

Table 1 Bayes Factor for One-Sample T-Test.

| Cumulative Cases | N    | Mean       | Std. Deviation | Std. Error Mean | Bayes Factors | t     | df  | Sig. (2-tailed) |
|------------------|------|------------|----------------|-----------------|---------------|-------|-----|----------------|
|                  | 680  | 94245.1574 | 26220.1592     | 1005.49721      | 0.000         | 93.730| 679 | 0.000          |

Bayes factor: Null versus alternative hypothesis.
Monte Carlo Sampling Seed: 1248774681.

Table 2 Posterior Distribution Description for One-Sample Mean.

| Cumulative Cases | N    | Mode | Mean       | Variance | 95% Credible Interval |
|------------------|------|------|------------|----------|-----------------------|
|                  | 680  |      | 94245.16   | 1017015.889 | 92267.98 - 96222.34   |
3.1. Expectation-Maximization Algorithm:

EM can be used for variables that cannot be directly observed and inferences can be drawn from the values of other observed variables [17,18]. It can be used with unnamed data to classify them. This is one of the common probability maximizations (to maximize the likelihood) methods.

3.2. Basic ideas of EM -Algorithm: Given a set of incomplete data and a set of starting parameters [19,20].

- **E-Step:** Using the given data and the current value of parameters, estimate the value of hidden data.
- **M-Step:** After the E-step, it is used to maximize the hidden variable and joint distribution of the data.
Repeat E-step and M-step until convergence.

4. The numerical results

Data sets are of the world health organization (WHO) were used from January 5, 2020, to 12, November 2021 [22], which represents a total of 677 days. As pointed by [23], although a country of interest is selected for the numerical study, the quantitative and qualitative conclusions are still valid. The selected time span was guided by our data preparation stage. We can suggest as future work to extend the time span to cover Omicron wave from December 2021 to February 2022 [24].

Table 3  Displays the returns brief of information for nodes in the predictive of the structure’s decision tree.

| Node | N   | Percent | Mean       |
|------|-----|---------|------------|
| 4    | 69  | 10.1%   | 110241.8551|
| 3    | 141 | 20.7%   | 101886.9929|
| 5    | 205 | 30.1%   | 98860.7073 |
| 2    | 128 | 18.8%   | 95712.1797 |
| 6    | 67  | 9.9%    | 72941.3284 |
| 1    | 70  | 10.3%   | 67275.4286 |

Table 4  Show the performance of the forecasting models.

|        | SMAPE | MSE    | MASE     |
|--------|-------|--------|----------|
| Nodes 6 | 0.102292182 | 103956.037 | 0.173279 |
| Nodes 5 | 0.10604745  | 107281.748 | 0.1784107|
| Nodes 3 | 0.102956911 | 109076.373 | 0.1761088|

![Evolution of the BIC for each model](image1)

**Fig. 5**  Evolution of the BIC for each model.

![Fitted model](image2)

**Fig. 6**  MAP Classification.

![Cumulative Distribution Function](image3)

**Fig. 8**  The empirical and estimated (The Cumulative Distribution Function).
4.1. Bayesian One Sample

See Tables 1 and 2.

4.2. Based on decision tree

See Figs. 3 and 4, Tables 3 and 4.

4.3. Based on GMM

Descriptive Statistics.
See Figs. 5-8.
Tables 5-8.

5. Discussion

Bayes Factor for One-Sample T-Test in Table 1. Show that the Null versus alternative hypothesis where the P-value is less than 0.005. Table 2 shows the Posterior Distribution Characterization for One-Sample Mean. Table 3. Displays the returns brief of information for nodes in the predictive of the structure’s decision tree. This indicates that the fifth node had the highest gains than the further nodes by 30.1%, which is indicative of the high level of effectiveness of the fifth observation factor in the predicting analysis of the cumulative cases. The performance of the forecasting models is shown in Table 4 where nodes 6 show higher efficiency than the rest of the nodes. Tables 5 and 6 show the Descriptive Statistics and evolu-

Table 5  Descriptive Statistics.

| Std. deviation | Mean       | Maximum          | Minimum | Obs. Without missing data | Obs. With missing data | Observations | Variable     |
|---------------|------------|------------------|--------|---------------------------|------------------------|--------------|--------------|
| 26198.398     | 94148.645  | 126836.000       | 0.000  | 678                       | 0                      | 678          | Cumulative Cases |

Table 6  Model classes.

| 5     | 4     | 3     | 2     | Model/Number of classes   |
|-------|-------|-------|-------|---------------------------|
| −15183.062 | −15316.774 | −15303.737 | −15744.35489 | E |

Table 7  The Proportions, the mean, and the variance by class.

| 5     | 4     | 3     | 2     | 1     | Proportions | Mean       | Variance   |
|-------|-------|-------|-------|-------|-------------|------------|------------|
| 120071.072 | 91723.359 | 91579.940 | 91556.305 | 9353.536 | 0.264       | 0.224      | 0.227      |
| 69157322.361 | 69157322.361 | 69157322.361 | 69157322.361 | 69157322.361 | 0.060       | 0.225      | 0.225      |

Table 8  The selection criterion for the selected model:

| DF | Entropy | NEC | Log-likelihood | ALL | AIC | BIC   |
|----|---------|-----|----------------|-----|-----|-------|
| 10.000 | 558.952   | 1.862      | −7558.935     | −16300.966 | −15137.871 | −15183.062 |

4.1. Bayesian One Sample

See Tables 1 and 2.

4.2. Based on decision tree

See Figs. 3 and 4, Tables 3 and 4.

4.3. Based on GMM

Descriptive Statistics.
See Figs. 5-8.
Tables 5-8.

5. Discussion

Bayes Factor for One-Sample T-Test in Table 1. Show that the Null versus alternative hypothesis where the P-value is less than 0.005. Table 2 shows the Posterior Distribution Characterization for One-Sample Mean. Table 3. Displays the returns brief of information for nodes in the predictive of the structure’s decision tree. This indicates that the fifth node had the highest gains than the further nodes by 30.1%, which is indicative of the high level of effectiveness of the fifth observation factor in the predicting analysis of the cumulative cases. The performance of the forecasting models is shown in Table 4 where nodes 6 show higher efficiency than the rest of the nodes. Tables 5 and 6 show the Descriptive Statistics and evo-

6. Conclusion

We have investigated presented an improvement of the usual methods of analysis, and prediction based on machine learning techniques, that has been realized by the application of MMG, DT approaches, tools for modeling, and evaluation of Knowledge of Covid-19 data in China. We have investigated enhance the use of the Gaussian mixture model and decision tree for the classification and prediction of new cases of coronavirus in China, both models showed efficiency in predicting and classifying new cases of coronavirus in China. GMM also can be used in different areas, such as marketing, finance, and so on. The models used in this study can be updated to include additional factors that affect the spread of the epidemic, such as health, demographics, and environmental factors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
Acknowledgements

This research project was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2022R125), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

References

[1] A.A. Elhag, T.A. Aloafi, T.M. Jawa, N. Sayed-Ahmed, F.S. Bayones, J. Bouwhi, Elhag et al, Artificial neural networks and statistical models for optimization studying COVID-19, Results Phys. 25 (2021) 104274, https://doi.org/10.1016/j.rinp.2021.104274.

[2] S. Bhardwaj, M. Alowaidi, R. Bhardwaj, S.K. Sharma, Machine learned hybrid Gaussian analysis of COVID-19 pandemic in India, Results Phys. 30 (2021) 104630, https://doi.org/10.1016/j.rinp.2021.104630.

[3] G. Perone, Comparison of ARIMA, ETS, NNAR, TBATS and hybrid models to forecast the second wave of COVID-19 hospitalizations in Italy, Eur. J. Health Econ. (2021).

[4] E. Gecili, A. Ziad, R.D. Szczesniak, A.M. Samy, Forecasting COVID-19 confirmed cases, deaths, and recoveries: Revisiting established time series modeling through novel applications for the USA and Italy, PloS One 16 (1) (2021) e0244173.

[5] I. Al-Turaiki, F. Almutlaq, H. Arafah, N. Alballa, Empirical Evaluation of Alternative Time-Series Models for COVID-19 Forecasting in Saudi Arabia, Int. J. Environ. Res. Public Health 18 (16) (2021) 8660, https://doi.org/10.3390/ijerph18168660.

[6] J.A. Borrego–Morell, E.J. Huertas, N. Torrado, C. Chen, On the effect of COVID-19 pandemic in the excess of human mortality. The case of Brazil and Spain, PLoS One 16 (9) (2021) e0255909, https://doi.org/10.1371/journal.pone.0255909.

[7] S.M. Abo-Dahab, M. Ragab, A.A. Elhag, S. Abdel-Khalek, et al, Free convection effect on oscillatory flow using artificial neural networks and statistical techniques, Alexandria Eng. J. 59 (5) (2020) 3599–3608.

[8] K. Alsabti, S. Ranka, V. Singh, CLOUDS: A Decision Tree Classifier for Large Datasets, in: Conference on Knowledge Discovery and Data Mining (KDD-98), August 1998.

[9] L. Zhao, S. Lee, S.-P. Jeong, Decision Tree Application to Classification Problems with Boosting Algorithm, Electronics 10 (16) (1903) 1903, https://doi.org/10.3390/electronics10161903.

[10] R. Sun, G. Wang, W. Zhang, L.-T. Hsu, W.-Y. Ochieng, A gradient boosting decision tree-based GPS signal reception classification algorithm, Appl. Soft Comput. 86 (2020) 105942, https://doi.org/10.1016/j.asoc.2019.105942.

[11] K.-C. Cheng, M.-J. Huang, C.-K. Fu, K.-H. Wang, H.-M. Wang, L.-H. Lin, Establishing a Multiple-Criteria Decision-Making Model for Stock Investment Decisions Using Data Mining Techniques, Sustainability 13 (6) (2021) 3100, https://doi.org/10.3390/su13063100.