MODELING AND CLASSIFICATION OF DEATHS DUE TO COVID-19 BASED ON MACHINE LEARNING TECHNIQUE

by

Randa ALHARBI*

Department of Statistics, Faculty of Science, University of Tabuk, Tabuk, Saudi Arabia

Original scientific paper
https://doi.org/10.2298/TSCI221015196A

Statistical classification is recently considered one of the most important and most common methods in machine learning models and consists of building models that define the target of research interest. There are many classification methods that can be used to predict the value of a response. In this article, we are interested in machine learning applications to classify the new deaths due to Covid-19. Under consideration BIC criterion, the experimental results have shown that the E (Equal variance) with four is the best mixture model. The convergence in the algorithm of expectation-maximization is satisfied after 167 iterations. The World Health Organization has presented the source of data over the period of March 2, 2020 to August 5, 2020.

Key words*: expectation-maximization algorithm, machine learning modeling, classification, Gaussian mixtures model, prediction*

Introduction

The prediction is carried out according to the selected numerical parameters using machine learning techniques. Machine learning is a mathematical discipline that allows, through the use of various sections of probability theory, mathematical statistics, and numerical methods, to obtain knowledge from available data. It is used to automate the solution of various tasks in a variety of areas of human activity. Nowadays, as a result of widespread informatization, impressive amounts of data have been accumulated in all kinds of industries such as manufacturing, science, business, and healthcare. To obtain promising results in the problem of the correct diagnosis of breast cancer doctors can be used various machine learning methods, [1].

Also, the different measuring can be employed to compare machine learning models. The degree of accuracy of machine learning models can be reached 98% other than other models. For the problem of comparison between Logistic Regression, SVM, Random Forest, and Naive Bayes the dataset of Wisconsin Breast cancer is used, [2]. The efficiency of the use, of algorithms is determined based on accuracy and time consumption in evaluating the correctness of classifying data. With the least error rate, we obtained that the Random Forest algorithm has the highest accuracy (99.76%) which is our main objective. The methodology

* Author's e-mail: ralharbi@ut.edu.sa

that is explained for the different three classification models SVM, DT, and ANN which are used to conduct the prediction are discussed.

In [3], the conclusion given in the paper for the experimental results is applied to test the validation of the models as well as sensitivity, accuracy, and specificity as criteria that are compared. In all the parameters of sensitivity, the results have shown that SVM performs well than Decision Tree and MLP. So, we can say that the best predictor is presented by SVM of breast cancer recurrence.

In [4], for the trend, a suitable prediction has been presented by a Gaussian mixture model which is developed in [5]. This model is used in a single day to ascertain the peak value and its data in a new case. In different parts of the world, the end-dates and number of cases of pandemic spread are compared with a similar study performed earlier and estimated by 95% confidence intervals. Also, with respect to [6], the machine learning library (MLlib) is used to implement the Gaussian mixture model under a spark environment.

For performing soft clustering, we use GMM model which probably model computes the probability of data points and put them in various Gaussians (clusters) [7]. The distributed dataset (RDD) is built-in by processing to mine useful data by distributing datasets by Spark performs parallel distributed processing. In memory, the scalability and computations are supported by apache spark. Hence, as clustering in GMM it works well for iterative algorithms [8].

Gaussian mixture model

Cluster operation of the multidimensional data by Gaussian mixture models according to distribution given by [9-11].

Mixture model

Let us assume that $x_1, x_2,\ldots,x_n$ are observed and x_i is a sample of K mixture component [11]. The mixture components were linked with random variables x_i where i =1,2,…,n are labeled as $Zi \in$ {1 to K} where it indicates which components x_i derive from.

The marginal probability of x_i taken from the law of total probability is given by:

$$P(X_i = x) = \sum_{k=1}^{k} (X_i = x \mid Z_i = k) P(Z_i = k) = \sum_{k=1}^{k} (X_i = x \mid Z_i = k) \pi_k \tag{1}$$

where the mixture weights or mixture proportions is defined by π_k and they the probability of x_i contained in the k^{th} mixture components. The mixture proportion is a non-negative and accumulated to 1, where $\Sigma_{k=1}^{k} \pi_k = 1$. Where $P(X_i \mid Z_i = k)$ called as the mixture component, and it characterizes the dispersal of x_i with the assumption that it has been derived from components k. The mixture component in this model under discussion is assumed to be normal distribution [13-15].

For random discrete variable, the mixture component can be described as mass functions probability (PMF), which can be expressed as $p(.\mid Z_k)$. For random continuous variable it can be described as density function probability (PDF), which can be expressed as $f(.\mid Z_k)$. [16].

The conforming PDF and PMF for the model of mixture can be written as:

$$p(x) = \sum_{k=1}^{k} \pi_k p(x \mid Z_k) \tag{2}$$

$$f_x(x) = \sum_{k=1}^{k} \pi_k f_x \mid z_k (x \mid Z_k) \tag{3}$$

If the independent discrete sample x_1, x_2,…,C are observed from the mixture, the mixture vectors proportion can be expressed as $\pi = (\pi_1, \pi_2,\ldots,\pi_k)$, so the function of likelihood can be written as:

$$L(\pi) = \prod_{i=1}^{n} P(X_i|\pi) = \prod_{i=1}^{n}\sum_{k=1}^{k} p(X_i \mid Z_i)\pi_k \tag{4}$$

Now assume we are in the mixture with Gaussian model and data setting with the component k^{th} can be modeled as $N(\mu_k,\sigma_k)$ with mixture proportion π_k.

Inference algorithm

Expectation maximization (EM) algorithm to estimate the parameters $\{\mu_k, \sigma_k, \pi_k\}$ under given observations x_1, x_2,…,x_n in the context Gaussian mixture models in [17,18]. Let, the random variable has normal probability distribution function $N(\mu,\sigma^2)$. Hence, the conditional distribution in this scenario given by:

$$X_k \mid Z_k = k \sim N\left(\mu_k, \sigma_k^2\right) \tag{5}$$

and therefore, the marginal probability of X_i given by:

$$P(X_i = x) = \sum_{k=1}^{k} P(Z_i = k) P(X_i = x \mid Z_i = k) = \sum_{k=1}^{k} \pi_k N\left(x; \mu_k, \sigma_k^2\right) \tag{6}$$

Hence, the joint case observations x_1, x_2,…,x_n is given by:

$$P(X_1 = x_1 \ldots\ldots\ldots X_1 = x_n) = \prod_{i=1}^{n}\sum_{k=1}^{k} \pi_k N\left(x_i; \mu_k, \sigma_k^2\right) \tag{7}$$

The eqs. (5)-(7) are described the Expectation maximization procedure with machine learning [18]. The goal is to achieve and estimate max-likelihood of π_k, μ_k, and σ_k^2 assumed that the database of observation $\{x_1, x_2,\ldots,x_n\}$.

Model selection

In a several cases, the mixture model contains an unknown number of components and several criteria such as Bayesian Information Criterion (BIC), [19, 20]. The BIC is applied information criterion which more similar to likelihood criterion penalized a model with several numbers of parameters [21]. The BIC model dependent on process of selected from group of candidate models under maximization the posterior probability [22].

Numerical results

The new deaths data under Covid-19 that available historical in Brazil over the period of March 2, 2020 to August 5, 2020. The data source is from World Health Organization [23].

Result and conclusion

In this study, we classified deaths in Brazil using the Gaussian mixtures model and discussed the enormous difficulty of predicting future deaths in Brazil. After 167 iterations, the EM algorithm's convergence was satisfied. As seen in tabs. 1 and 2, the mixture model with four components and equal variance is the optimal model, according to the Bayesian information criteria. The proportions of the four components range from 0.1 to 0.3, tab. 3. The range of the four components' means is 71 to 1174, tab. 4. Four components' variance is 16786.282, tab. 5. Model selection using the chosen criterion. The NEC, which indicates that

there is a clustering structure in the data, is the smallest requirement, tab. 6. From these results, it can be said that the Gaussian mixture models showed a high predictive efficiency, so it should be noted that the implementation of such a mechanism to predict the new deaths data under Covid-19 is extremely useful, figs. 1-5.

Table 1. The descriptive statistics for new deaths due to Covid-19 in Brazil over the period of March 2, 2020 to August 5, 2020, for 162 observations (Obs.)

Statistics results							
S.D	Mean	Maximum	Minimum	Obs. under missing data	Obs. under missing data	Obs.	Variable
487.212	584.352	1595	0	162	0	162	New deaths

Table 2. The BIC selections

BIC computation under each model				
Model/Number of classes	2	3	4	5
E	–2437.364	–2447.539	–2403.158	–2413.333

Table 3. Different four components and its proportions

Proportions				
Class	1	2	3	4
Proportions	0.397	0.275	0.164	0.164

Table.4 The mean by the four components

Means by class				
Class	1	2	3	4
Mean (New deaths)	70.853	623.806	1174.290	1174.290

Table 5. The variance by the four components

Variance by class				
Class	1	2	3	4
Variance	16786.282	16786.282	16786.282	16786.282

Table 6. The model selection under selected criterion; the smallest one criterion is NEC which indicate that there is a clustering structure in the data

The model selections and its criterion						
DF	Entropy	NEC	Log-likelihood	ICL	AIC	BIC
8.000	49.557	0.977	–1181.229	–2502.272	–2378.457	–2403.158

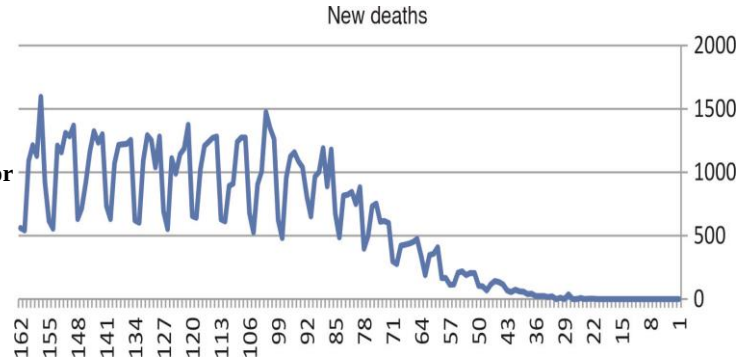

Figure 1. Shows the series for new deaths due to Covid-19 Brazil over the period of March 2, 2020 to August 5, 2020

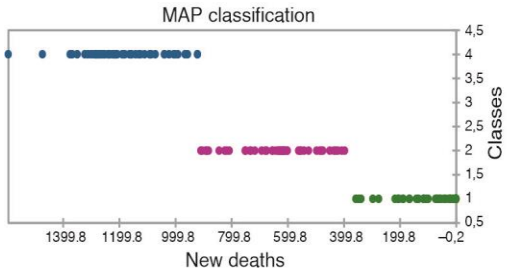

Figure 2. No assignment exists in Classes 3 by the MAP classification

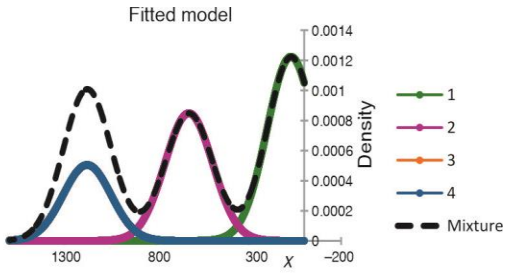

Figure 3. Shows the mixture model

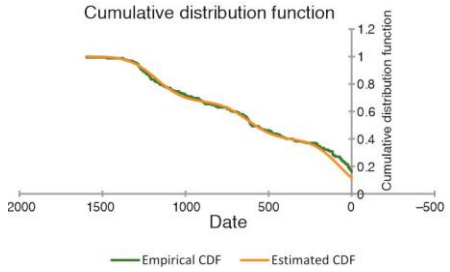

Figure 4. The estimated CDF and empirical CDF are very close under the mixture model, which satisfies the accuracy of the estimation

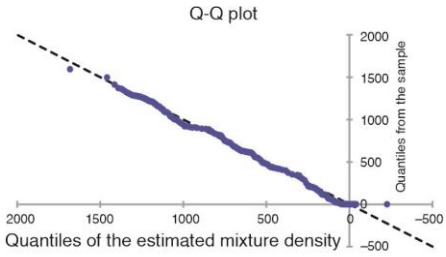

Figure 5. Shows the quintiles of the estimated mixture density

References

[1] Janghel, R. R., *et al*., Classification and Detection of Breast Cancer Using Machine Learning, *Social Networking and Computational Intelligence*, *100* (2020), Mar., pp. 269-282
[2] Sivapriya, J., *et al*., Breast Cancer Prediction using Machine Learning, *International Journal of Recent Technology and Engineering (IJRTE) 8* (2019), 4, pp. 2277-3878
[3] Elmustafa, S. A., *et al*., Machine Learning Technologies for Secure Vehicular Communication in Internet of Vehicles: Recent Advances and Applications, *Journal of Security and Communication Networks (SCN), 2021* (2021), ID8868355
[4] Ahmad, L. G., *et al*., Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence, *J. Health Med. Inform., 4* (2013), 124
[5] Singhala, A., Modeling and Prediction of COVID-19 Pandemic Using Gaussian Mixture, *Chaos Solitons & Fractals, 138* (2020), 110023
[6] Lavanya, K., *et al*., Clustering of Zika Virus Epidemic Using Gaussian Mixture Model in Spark Environment, *Biomedical Research, 30* (2019), Jan., pp. 127-133
[7] ***, https://www.kaggle.com/vipulgandhi/gaussian-mixture-models-clustering-explained
[8] Sarkar, S., *et al*., Gaussian Mixture Modeling and Model-Based Clustering Under Measurement Inconsistency, *Adv Data Anal Classif, 14* (2020), May, pp. 379-413
[9] Eva, P., *et al*., Clustering Cloud Workloads: K-Means *vs*. Gaussian Mixture Model, *Procedia Computer Science, 171* (2020), Jan., pp. 158-167
[10] Wamba, G. M., *et al*., Cloud Workload Prediction and Generation Models, *Proceedings*, 29th International Symposium on Computer Architecture and High-Performance Computing, Campinas, Brazil, 2017,pp. 89-96
[11] Rayan, A. A., *et al*., Analysis and Challenges of Robust E-Exams Performance Under COVID-19, *Results in Physics, 23* (2021), 103987
[12] Li, Y., *et al*., A Gaussian Mixture Model to Detect Clusters Embedded in Feature Subspace, *Communications in Information System, 7* (2007), 4, pp. 337-352

[13] Constantinopoulos, M. K., *et al*., Bayesian Feature and Model Selection for Gaussian Mixture Models, *IEEE Trans. on PAMI, 28* (2006), 6, pp. 1013-1018
[14] Hassan, M. B., *et al*., Machine Learning for Industrial IoT Systems, in: *Handbook of Research on Innovations and Applications of Al, IoT, and Cognitive Technologies*, IGI Global, Hershey, Penn., USA, 2021
[15] Kumar, V. V., *Applications of AI, IoT, and Cognitive Technologies*, (ed. Zhao, J.,) Hershey, Penn., USA, 2021, pp. 336-358
[16] Salih, A., *et al*., Machine Learning in Cyber-Physical Systems in Industry 4.0, in: *Artificial Intelligence Paradigms for Smart Cyber-Physical Systems*, (eds. Luhach, A. K., Atilla, E.,) Hershey, Penn., USA, 2021, pp. 20-41
[17] Azhari, A. E., Almarashi, A. M., Forecasting Based on Some Statistical and Machine Learning Methods, *Journal of Information Science and Engineering, 36* (2020), 6, pp. 1167-1177
[18] Liu, X., *et al*., Gaussian Mixture Models Clustering Using Markov Random Field for Multispectral Remote Sensing Images, *Proceedings*, International Conference on Machine Learning and Cybernetics, Dalian, China, 2006, pp. 4155-4159
[19] Verbeek, J. J., *et al*., Efficient Greedy Learning of Gaussian Mixture Models, *Published in Neural Computation, 15* (2003), 2, pp. 469-485
[20] Cron, A. J., West, M. Efficient Classification-Based Relabeling in Mixture Models, *The American Statistician, 65* (2011), 1, pp. 16-20
[21] Alsharif, S, *et al*., An Efficient HAPS Cross-Layer Design to Mitigate COVID-19 Consequences, *Intelligent Automation & Soft Computing, 31* (2022), 1, pp. 43-59
[22] Aljohani, H. M., Elhag, A. A., Using Statistical Model to Study the Daily Closing Price Index in the Kingdom of Saudi Arabia (KSA), *Complexity, 2021* (2021), ID5593273
[23] ***, https://www.who.int/emergencies/diseases/novel-coronavirus-2019

Paper submitted: October 15, 2022
Paper revised: November 20, 2022
Paper accepted: November 26, 2022