A Comparative Study of Classification Methods on the States of the USA Based on COVID-19 Indicators

İbrahim Miraç Eligüzel, Eren Özceylan

A Comparative Study of Classification Methods on the States of the USA based on COVID-19 Indicators

İbrahim Miraç Eligüzel[0000-0003-3105-9438] and Eren Özceylan[0000-0002-5213-6335]

Industrial Engineering Department, Gaziantep University, 27100, Gaziantep, Turkey
miraceliguzel@gmail.com, erenozceylan@gmail.com

Abstract. COVID-19 spreads across the world and specific pre-caution strategies are required for different regions depending on the current satiation. Therefore, proper region-specific pre-caution processes occupy a significant place to tackle with COVID-19 pandemic. During the COVID-19 pandemic, significant data are cumulated and these data can be utilized in order to cope with pandemic efficiently via providing a better understanding for decision-makers. In the aforementioned aspect, data related to the COVID-19 pandemic is used to decide on group states where the application of the same pre-caution processes has become efficient and effective. Therefore, COVID-19 indicators (e.g. number of deaths and infected) can be utilized to cluster the states, regions, countries, etc. In order to accomplish the underlined objective, data with seven features (rate of one dose, rate of two doses, number of cases, death, tests, recovered people, and percentage of positive tests) are retrieved for each of 50 states in the USA. After that, a dissimilarity matrix for cities is generated with respect to the corresponding seven features. Lastly, clustering methods (K-means, Agglomerative Hierarchical and BIRCH clustering, and P-median model) in literature are applied to gather clusters of states. In the proposed study, 50 states are taken into consideration and four different methods are applied to divide states into 6 subsets. The best result is gathered via the K-means application.

Keywords: Agglomerative Hierarchical clustering, BIRCH clustering, COVID-19, K-means clustering, P-median.

1 Introduction

With the spread of COVID-19 through the world, authorities required to take necessary pre-cautions to avoid the public health crisis. In the world most of the countries, regions and states demonstrate the different trends of COVID-19 spread, number of deaths, cases and recoveries. Therefore, it is necessary to adopt differentiated procedures for different trends. In addition, quick response is required to handle COVID-19 with respect to trends and it requires efficient and effective decision making process. In this study our aim is to demonstrate applicability of clustering algorithms on problems in which, geographical classification occupies a crucial place. In addition, focus of this study is to provide scientific approach for decision makers against the COVID-19 pandemic. Therefore, decision makers can determine the policies and apply them in accordance with the proposed geographical clusters.

Decisions comprehend travel and mobility restriction, social distancing and quarantine requirements, school and business closures, large scale testing, forbidding large crowd gatherings[1]. Moreover, there are organizations that generate indexes for countries and cluster them in accordance. For instance, Deep Knowledge Group conducted a study that focuses on COVID-19 regional safety assessment [2] and INFORM (Information Network Focus on Religious Movements) provides a report related to risk index for countries [3]. Results and methods of this study can be utilized in order to generate indexes as in explained above. There are several studies conducted to have deeper insight to COVID-19 pandemic due to developing strategies against it. In this manner, the study is conducted by using the fuzzy clustering for countries and Pearson correlation is applied find meaningful relation between the data features, in which strong correlation is detected between confirmed cases, dead cases and population size[4]. Another study that aims to support decision making process propose novel clustering algorithm adapted to make comparison of the various COVID time-series of different countries [5]. From the aforementioned perspective, it is also important to analyze spread of the COVID-19. Therefore, spatiotemporal distribution assessment and spread pattern analysis of COVID-19 is done by utilizing clustering algorithm and geographical information system [6]. Early detection of COVID-19 can also be thought as a crucial step of handling the pandemic. There are several ways to detect COVID-19. Chest X-ray CT-scan images are rather cheaper way to implement. In this aspect, a novel meta-heuristic feature selection method (Golden Ratio Optimizer) is proposed to extract features from images and it demonstrates outstanding performance with accuracies more than 99% [7]. Another study focuses on time series analysis which occupies a significant place to extract the dissimilarity in the COVID-19 spread in states and the establishment of models which can contribute analysis and prediction for transmission process of this infectious disease [8]. Relaxation of the restriction for COVID-19 is also important concern for economic and social life. The study is conducted to develop strategy to apply gradual relaxation by analyzing an epidemiological compartmental model with multi-objective genetic algorithm design optimization in order to make comparison on scenarios of strategy type [9]. As it is stated above, better understanding of COVID-19 spread is crucial to tackle pandemic. In this manner study is conducted to detect relation between living environment deprivation and spatial clustering of COVID-19 hotspots, and results show that living environment deprivation is significant determinant [10]. Social impact of the COVID-19 is taken into consideration during the pandemic. The best way to analyze the social impact is utilizing the social media. One of the studies uses Twitter hash-tags and analyze them with multi-view clustering method and results demonstrates that some topical cluster of hash-tags stay consistent, while some of them shift during the pandemic [11]. Deciding on right protocol is a crucial step for the cases which require precision planning. Ramirez-Nafarrate et. al. conducted a study that includes the optimal planning for emergency departments to utilize ambulance diversion to ease congestion by requesting that ambulances bypass the ED and move patients to another location [12]. All in all, it can be concluded that having a deep insight of COVID-19 and application of right decisions for right place is very crucial to handle COVID-19 pandemic. In this study, it is aimed to gather

clusters of states in USA to apply same pre-caution strategies on states in same cluster. Moreover, agglomerative hierarchical clustering, BIRCH clustering, K-means clustering and P-median are applied on the data-set that retrieved from Johns Hopkins Coronavirus Resource Center [13], The New York Times [14] and worldometer [15]. In the combined data-set seven features are gathered and used for generating dissimilarity matrix based on Euclidean distance. After that aforementioned methods applied and performance evaluation each method is done by total distances of cluster from its centroid. Therefore, efficient and effective application of strategies can be implemented and general view of current situation can be provided for decision makers. Also, proposed study differentiated from other studies by considering vaccine dose data in the dataset. The rest of the study is organized as follows: Section 2 introduces the methods; the experiments are given in Section 3. Section 4 presents the conclusion.

2 Methodology

Four methods are applied on retrieved data-set in order to cluster given 50 states with respect to data with seven features which are rate of one dose, rate of two doses, number of cases, number of deaths, number of tests, number of recovered people, and percentage of positive tests. After that Euclidean distance is calculated for each pair of feature to gather dissimilarity matrix. Then, agglomerative hierarchical clustering, BIRCH clustering, K-means clustering and P-median methods applied in order to gather group of states. Number of cluster is decided by elbow graph via K-means application and six clusters are generated by application of each method. There is also Silhoutte score to find the right number of clusters. Since, elbow method gives the result that close the optimal number of clusters [16], it is utilized in this study. Another reason to utilize elbow method is to identify K beyond which the inertia doesn't change much.

2.1 K-means Clustering

The algorithm works with a set of *d*-dimensional (feature) vectors, $D = \{x_i \mid i = 1,..., N\}$, where $x_i^{(k)}$ denotes the i^{th} point and initialization of algorithm starts with selecting random *k* points in *d* as centroids to set them as the solution of clustering [17]. Also, c_j denotes center of cluster *j*. Objective function of K-means is given in Eq. (1).

$$\sum_{j=1}^{k} \sum_{i=1}^{n} \left(xi^{(k)} - cj \right)^2 \qquad (1)$$

In order to apply K-means, "sklearn" library is used. In this library, K-means is applied with giving number of clusters, initialization procedure, maximum number iteration and randomness procedure. Number of clusters is decided by applying elbow method (see Fig. 1) and number of cluster is decided as 6 for four algorithms.

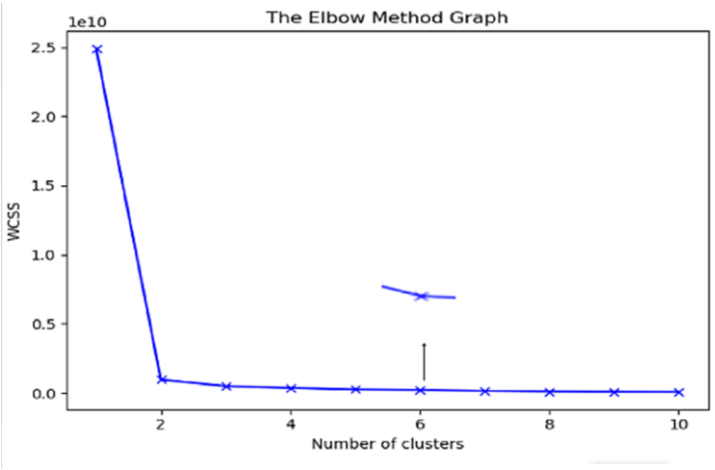

Fig. 1. Elbow method graph for deciding number of clusters

After the application of K-means, clusters are gathered and centroids of each cluster are found. Total distance for each cluster calculated with respect to Euclidean distance and total distances are summed in order to demonstrate the general performance of K-means.

2.2 Agglomerative Hierarchical Clustering

The single linkage base method for agglomerative hierarchical clustering algorithm and the process of this algorithm consists of initialization which regards each point as a single cluster. Then, connection of a pair of clusters is implemented by aiming nearest members (states) with shortest distance and this procedure is repeated till only one cluster remains [18]. In addition, "ward" method is utilized in this algorithm. "Ward" method aims to minimize the variance for each cluster. Its difference from K-means is instead of minimizing within cluster sum of square, it focuses on minimizing the within-cluster variants. In order to perform this algorithm, "scipy" and "sklearn" Python libraries are utilized together.

2.3 BIRCH Clustering

BIRCH is an annotation for balanced iterative reducing and clustering using hierarchies and it is an unsupervised method which is grouping the given dataset into clusters. For this method "sklearn" Python library is utilized. BIRCH parameters can be expressed as the branching factor *Br*, the threshold *T*, and the cluster count *k* and the steps are given as follows [19]:

- When data is given to BIRCH, one of the trees which are a height-balanced, the cluster features, and *CF* of is generated.

- Each node indicates a subset in the cluster hierarchy, intermediate nodes present super-clusters and the leaf nodes are considered as actual clusters.
- The branching factor *Br* is the maximum number of offspring that allowable for a one which is a global parameter. Every node comprehends the information of the cluster which is the cluster features (*CF*). The cluster centers (C_i) and cluster radius (R_i) can be calculated, where $\{x_{ij}\}_{j=1}^{n}$ are the elements of the i[th] cluster. Formulas for C_i and R_i given in Eqs. (2) and (3), respectively.

$$R_i = \sqrt{{}^{1}/_{n_i} \sum_j^n (x_{ij} - C_i)^2} \quad (2)$$

$$C_i = {}^{1}/_{n_i} \sum_j^n x_{ij} \quad (3)$$

- Every new point begins with root and recursively proceeds the tree, until the walk ends at a leaf node with nearest center.
- By arriving a leaf, the new point is included in given leaf cluster with respect to radius of the cluster threshold *T*. By exceeding *T*, a new cluster is generated with the new point. Therefore, the threshold parameter controls the size of the clusters.

By evaluation method which is used for two other aforementioned algorithms is applied.

2.4 P-median Method

The P-median problem and its variations has its crucial place to be applied on real life situations [19]. Considering a set L of *m* facilities, a set *U* of *n* users, and *n*x*m* matrix *D* indicates distances between two points d_{ij} in which *i* is demand from the facility *j*, for all $j \in L$ and $i \in U$ [19]. The objective function and variables are given below.

$$y_j = \begin{cases} 1, if\ a\ center\ state\ is\ opened\ in\ j \in L, \\ 0, otherwise \end{cases}$$

$$x_{ij} = \begin{cases} 1, if\ state\ i\ is\ attended\ \text{to}\ center\ state\ j \in L, \\ 0, otherwise \end{cases}$$

$$min \sum_i \sum_j d_{ij} x_{ij} \quad (4)$$

Subject to

$$\sum_j x_{ij} = 1, \forall_i, \quad (5)$$

$$x_{ij} \leq y_j, \forall_{i,j}, \quad (6)$$

$$\sum_j y_j = p, \quad (7)$$

$$x_{ij}, y_j \in \{0, 1\}. \quad (8)$$

The objective function (Eq. 4) aims to minimize total distance. In P-median model each state must be attended to only one subset (Eq. 5). Eq. (6) ensures that if a center state is opened then a state can be assigned to that center. Eq. (7) provides number of clusters do not exceed the pre-decided number of clusters. The signs of the decision variables are shown in Eq. (8). After applying P-median, centroid of each of six subsets are gathered by using the "numpy" Python library. Then, same evaluation with rest of the aforementioned algorithms is applied.

3 Clustering of the States

Utilized dataset consists of the seven features and features for is state is treated as dimensional vector to calculate dissimilarities. The illustration of dataset for states is given in Table 1. 1st and 2nd features are retrieved from The New York Times [14], 3rd, 4rd, 5th and 7th features are gathered from Johns Hopkins Coronavirus Resource Center [13] and 6th feature is taken from worldometer [15].

Table 1. Used dataset

#	State	Rate of one dose (in population)	Rate of two doses (in population)	Cases per 100k population	Deaths per 100k population	Tests per 100k population	Recovered people per 100k population	Percentage of positive tests
1	Alabama	10%	3.40%	9,879	192	93.30	5,613.60	21,22%
2	Arizona	13%	4.00%	11,034	210	394.40	1,556.19	5,52%
3	Arkansas	11%	4.70%	10,420	176	150.20	9,999.72	10.72%
…	…	…	…	…	…	…	…	…
48	West Virginia	14%	8.30%	7,185	125	493.90	6,497.01	3,78%
49	Wisconsin	13%	4.60%	10,478	117	429.70	9,300.49	3,07%
50	Wyoming	13%	5.20%	9,249	114	403.90	9,010.49	3,37%

For making comparison between aforementioned models and clustering algorithms, median of each cluster is found and total distance for each cluster is calculated with respect to medians to receive aggregate distance for each application via Euclidean distance. Therefore, evaluation of each application is implemented. Meanwhile, Python 3.7 libraries are utilized with a computer that has Intel Core i7 9750H CPU with 8GB RAM features. In this study, total distances of applications which are K-means, agglomerative hierarchical clustering, BIRCH clustering and P-median are gathered as 82,914.43, 89,597.34, 89,597.34 and 85,225.59 respectively. Total distances are calculated with given equation below (Eq. 9).

$$\sqrt{\sum_j^k \sum_i^n \left(x_{ji} - c_j\right)^2} \qquad (9)$$

Where x_{ji} indicates the data point "i" that attended to cluster "j" and c_j indicates the center point of cluster "j".

All in all, K-means shows better result compared to others, which is followed by P-median method. In addition, Agglomerative hierarchical clustering and BIRCH clus-

tering show exactly the same performance. In nutshell, purpose of the study is to cluster the states in order to apply same procedure for each cluster of states against pandemic. Six clusters are gathered from aforementioned four methods. The best result is gathered from K-means application. Demonstrations of the results obtained by four methods are given in Fig. 2.

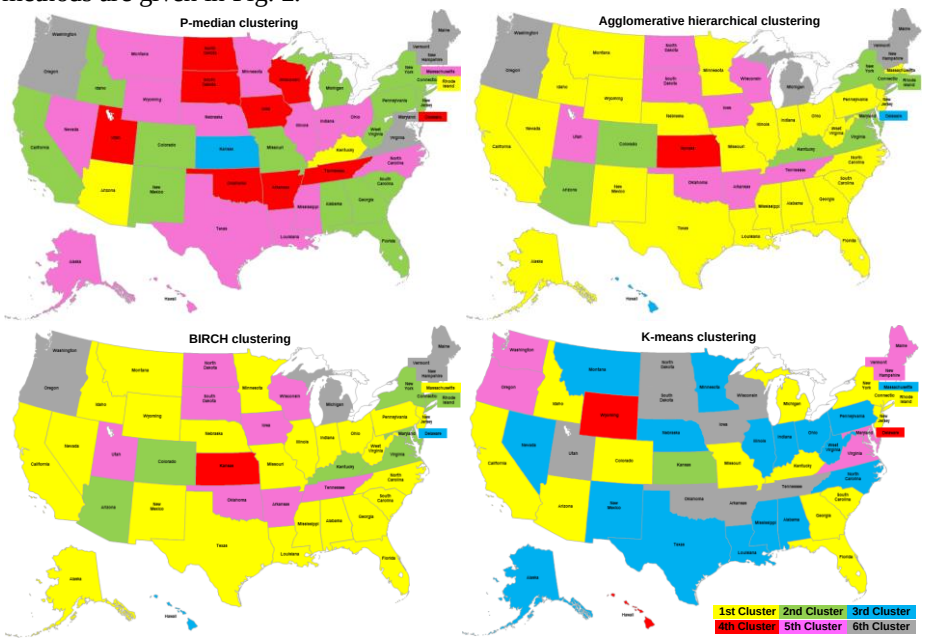

Fig. 2. The clustered states obtained by four methods

Clusters are represented by different colors in Fig.2. Each cluster indicates the collection of states, where same strategic decisions can be applied against COVID-19 pandemic. For instance, cities significantly differ from each other by "percentage of positive tests", "tests per 100k population" and "rate of one dose" in 4th cluster and 2nd cluster for K-means application. However, there is no significant or consistent difference for rest of the attributes for cities. From the Fig. 2 it can be seen that Arkansas, Illinois, North Dakota, Oklahoma, South Dakota, Tennessee, Utah, and Wisconsin are assigned in the same cluster in all methods. Therefore, it can be concluded that same strategy against COVID-19 can be applied whole aforementioned eight states.

4 Conclusion

The main focus of study is to implement four clustering methods on USA states in order to cluster them with respect to COVID-19 via using extracted dissimilarity matrix from dataset with seven features. As a result, total distances of four applications (K-means, agglomerative hierarchical clustering, BIRCH clustering and P-median) are calculated as 82,914.43, 89,597.34, 89,597.34 and 85,225.59 in accordance. K-

means gives the better results compared to rest of the applications. Benefit of proposed study is providing information for decision makers to have general view of current situation geographically by aiming application of specific protocol on a gathered cluster. Limitations of this study can be expressed as required time to obtain data and evaluating several different features to analyze their effects on clusters. For the feature work, effect of each feature on the result can be investigated. Therefore, effectiveness of application can be augmented. For the future study, a web-based dynamic system can be developed due to being updated because COVID-19 data is altering over time. Lastly, utilized methods can be improvement with some modifications to increase performance. This study conducted on COVID-19 data in order to demonstrate geographical relation in USA states to provide inside for decision makers. In aforementioned aspect, given approach can be generalized for the similar situations with cumulated data. In addition, it can be used for classifications of geographic regions for different concerns to develop strategies, such as crime, politics and other pandemics situations.

References

1. G. M. Campedelli and M. R. D'Orsogna, "Temporal Clustering of Disorder Events During the COVID-19 Pandemic," pp. 1–32, arXiv preprint arXiv:2101.06458, (2021).
2. Deep Knowledge Group, http://analytics.dkv.global/covid-regional-assessment-200-regions/infographic-summary.pdf, last accessed 2021/03/16.
3. INFORM,https://reliefweb.int/sites/reliefweb.int/files/resources/INFORM%20COVID%20Risk%20Index%20V012%20Report.pdf, last accessed 2021/03/16.
4. M. R. Mahmoudi, D. Baleanu, Z. Mansor, B. A. Tuan, and K. H. Pho, "Fuzzy clustering method to compare the spread rate of Covid-19 in the high risks countries," Chaos, Solitons and Fractals, vol. 140, pp. 1–9, (2020).
5. V. Zarikas, S. G. Poulopoulos, Z. Gareiou, and E. Zervas, "Clustering analysis of countries using the COVID-19 cases dataset," Data in Brief, vol. 31, pp. 105787, (2020).
6. M. Azarafza, M. Azarafza, and H. Akgün, "Clustering method for spread pattern analysis of corona-virus (COVID-19) infection in Iran," medRxiv, (2020).
7. S. Chattopadhyay, A. Dey, P. K. Singh, Z. W. Geem, and R. Sarkar, "Covid-19 Detection by Optimizing Deep Residual Features with Improved Clustering-Based Golden Ratio Optimizer," Diagnostics, vol. 11, no. 2, pp. 315, (2021).
8. F. Rojas, O. Valenzuela, and I. Rojas, "Estimation of COVID-19 dynamics in the different states of the United States using time-series clustering," medRxiv, (2020).
9. O. Pinto Neto et al., "Mathematical model of COVID-19 intervention scenarios for São Paulo—Brazil," Nature Communications, vol. 12, no. 1, pp. 1–13, (2021).
10. A. Das, S. Ghosh, K. Das, T. Basu, I. Dutta, and M. Das, "Living environment matters: Unravelling the spatial clustering of COVID-19 hotspots in Kolkata megacity, India," Sustainable Cities and Society, vol. 65, no. October 2020, pp. 102577, (2021).
11. I. J. Cruickshank and K. M. Carley, "Characterizing Communities of Hashtag Usage on Twitter during the 2020 COVID-19 Pandemic by Multi-view Clustering", Applied Network Science, vol. 5, no. 66, pp. 1–40, (2020).

12. Ramirez-Nafarrate, A., Hafizoglu, A. B., Gel, E. S., & Fowler, J. W. (2014). Optimal control policies for ambulance diversion. *European Journal of Operational Research*, *236*(1), 298-312.
13. Johns Hopkins University, https://coronavirus.jhu.edu/testing/states-comparison/testing-state-totals-bypop, last accessed 2021/03/16.
14. New York Times, https://www.nytimes.com/interactive/2020/us/covid-19-vaccine-doses.html, last accessed 2021/03/16.
15. Report Coronavirus Cases, https://www.worldometers.info/coronavirus/country/us/, last accessed 2021/03/16.
16. Et-taleby, A., Boussetta, M., & Benslimane, M. Faults Detection for Photovoltaic Field Based on K-Means, Elbow, and Average Silhouette Techniques through the Segmentation of a Thermal Image. International Journal of Photoenergy, (2020).
17. X. Wu et al., Top 10 algorithms in data mining, Knowledge and Information Systems, vol. 14, pp. 1–37, (2008).
18. H. Koga, T. Ishibashi, and T. Watanabe, "Fast agglomerative hierarchical clustering algorithm using locality-sensitive hashing," Knowl. Inf. Syst., vol. 12, no. 1, pp. 25–53, (2007).
19. B. Lorbeer, A. Kosareva, B. Deva, D. Softić, P. Ruppel, and A. Küpper, "Variations on the Clustering Algorithm BIRCH," Big Data Res., vol. 11, pp. 44–53, (2018).
20. M. Dzator and J. Dzator, "An effective heuristic for the P-median problem with application to ambulance location," Opsearch, vol. 50, no. 1, pp. 60–74, (2013).
21. N. Mladenović, J. Brimberg, P. Hansen, and J. A. Moreno-Pérez, "The p-median problem: A survey of metaheuristic approaches," Eur. J. Oper. Res., vol. 179, no. 3, pp. 927–939, (2007).