2021 20th International Symposium INFOTEH-JAHORINA (INFOTEH) | 978-1-7281-8229-2/20/$31.00 ©2021 IEEE | DOI: 10.1109/INFOTEH51037.2021.9400670

Clustering of European countries and territories based on cumulative relative number of COVID 19 patients in 2020

Vukašin Crnogorac, Milana Grbić, Marko Đukanović, Dragan Matić
Dept. for Mathematics and Informatics, Faculty of Natural Science and Mathematics,
Banja Luka, Bosnia and Herzegovina
vukasin.crnogorac@pmf.unibl.org
milana.grbic@pmf.unibl.org
marko.djukanovic@pmf.unibl.org
dragan.matic@pmf.unibl.org

Abstract— **The coronavirus COVID-19 has been affected all the countries and territories in 2020. In this study we cluster European countries according to the cumulative relative number of European COVID-19 patients. The clustering is based on publicly available data published at European Centre for Disease Prevention and Control website and performed by three clustering methods: K-means, agglomerative and BIRCH clustering. Clustering performance, evaluated by Silhouette Coefficient value, shows satisfying accuracy of the obtained clusters. The results presented in this study can be useful to public health officers and practitioners to easier deal with COVID-19 challenges.**

Keywords: K-means clustering; Agglomerative clustering; BIRCH clustering; COVID-19; European countries

I. Introduction

The coronavirus COVID-19 has been affected by more than 200 countries and territories in 2020. After the coronavirus firstly appeared in China at the end of 2019., it promptly became a global threat for the entire world population. The first individual cases in Europe have been registered at the beginning of 2020. Although most of the countries introduced strong measures to tackle rising infection rates, the coronavirus was spread quickly from one country to another. Novel cases increased ten-fold in monthly level: from 100,000 in the first week of March to more than one million on 02 April, with more than 52,000 deaths reported across the world. Unfortunately, such a trend has continued to the present day and in the time when this research is made (beginning of 2021.), more than 86 million coronavirus cases are registered and more than 1,800,000 deaths are reported by World Health Organization (WHO), (see https://covid19.who.int/).

Many researchers from various scientific fields are working together to produce relevant information on COVID-19 risk factors to slow down the spread of the disease and to reduce the number of severe cases and deaths. In this line, computer scientists are focused on analyzing the existing statistical data, based on both individual (like patients and their characteristics) or global levels. Machine learning techniques may help to identify trends and patterns which could help public health officers and practitioners to easier deal with this challenging threat of the modern age.

In this paper, we contribute by analyzing the cumulative number for 14 days of COVID-19 cases per 100000 over the period of about 9 months (from April to the mid-December 2020.) in European countries and territories. The aim of our research is to group European countries and territories in such groups/clusters, where countries from a cluster have similar values of the cumulative number of COVID-19 cases. This research could help public health officers and other authorities to better plan their politics regarding the pandemic.

II. Previous work

From the beginning of COVID-19 pandemic, on a daily base, a lot of data is generated. There are several types of researches which are dealing with different types of data associated with COVID-19 pandemic.

A comprehensive survey of artificial intelligence approaches in tackling the pandemic problem in many aspects is given in [1]. The aforementioned paper is organized to cover the following applications: clinical applications, processing COVID-19 related images, pharmaceutical studies and epidemiology. For each application, the research is subdivided based on the artificial intelligence approaches they have employed. The survey which covers both medical and technological perspectives to facilitate the virologists, artificial intelligence researchers and policymakers while in combating the COVID-19 outbreak is given in [2]. A specific aspect of the pandemic on increasing test capacity via the optimal allocation of swabs and reagents to laboratories is processed by proposing an Integer Programming model in [3]. A mathematical program to model the aspect of establishing how many diagnostic tests the regions (in Italy) have to perform to maximize the overall disease detection capability is proposed in [4]. A stochastic optimization model for allocating and sharing a critical resource in the case of a COVID-19 pandemic is considered in [5].

Beside surveys and theoretical researches, various applications have been developed during the pandemic to facilitate days activities and decrease the possibility of infection. In [6], a machine learning-based prognostic model and mobile application software platform for predicting infection susceptibility of COVID-19 using healthcare data are proposed.

Total number of papers dealing with clustering of COVID-19 data is increasing on daily basis, so a detailed survey of such applications is out of the scope of this paper. Here we mention only a few related results. A clustering model, based on open-access variables at the country level, that can group countries well regarding the number of confirmed COVID-19 cases was developed in [7]. Similarly, in [8] the hidden community structure of countries by applying the unsupervised clustering approach based on the trend, countries affected regionally and the variety of cases at the country level on COVID-19 dataset, are detected. Analyzing time-series of COVID data, such as active cases, active cases per population and active cases per population and per area for different countries is performed in [9]. In the same paper, new specially designed clustering algorithm is proposed. Similarly, cluster-based method which partition countries on daily basis, analyzes the evolution of multivariate time series and applies this to the COVID-19 pandemic data [10]. The proposed method demonstrates a close similarity in the evolution of cases and deaths. In [11] the correlation between spread of Covid-19 and population's size is studied, after that the fuzzy clustering is applied on countries with high spread risk.

Comparing to other approaches, in our paper we propose a different approach for clustering European countries, which is based on the cumulative relative number of COVID-19 cases in Europe. Thus, we start from a different starting point and impose the three different methods for clustering the data.

III. The Method

In this section we explain the aim of our work and describe the proposed methods for clustering the European countries and territories based on biweekly growth the number of confirmed COVID-19 cases.

A. Aim of work

Our task is grouping (clustering) European countries and territories with a similar 14-day cumulative number of reported COVID-19 cases per 100 000 population. To achieve that, we use three different clustering techniques, namely, K-means, agglomerative and Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) clustering. Some of these techniques are also used in health and clinical research [12-14]. We further analyzed obtained clusters and discussed them concerning the position of a country in a region.

B. Data

In this research we used the original raw data about new COVID-19 cases on daily basis, published at European Centre for Disease Prevention and Control (ECDPC) website (available at: https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide). Data can be downloaded in various formats: .xlsx, .json, .xml and .csv formats. For purpose of this research, we used data in .xlsx format, where each row/entry contains the number of new cases reported per day and per country with additional information about the 14-day cumulative number of reported COVID-19 cases per 100 000 population in 2020. According to the information presented at ECDPC website, it is noted that: (i) Population data in the database is taken from Eurostat for Europe and the World Bank for the rest of the world. (ii) Countries that are not listed in these databases have reported no cases to WHO and no cases were identified in the public domain.

Since COVID-19 appeared in European countries in different periods at the beginning of 2020., there are some missing data for the period of January – March 2020. Thus, in our research, we observed the period from 14th April to 8th December 2020. We chose this period since the pandemic of COVID-19 has been detected in March in most of European countries.

We didn't consider absolute numbers of confirmed COVID-19 cases, since there is a huge difference in population size and also in number of tested cases among countries. For that reason, we took into account the information about 14-day cumulative number of reported COVID-19 cases per 100 000 population. More precisely, we extracted the cumulative number of reported COVID-19 cases per 100 000 population, which is calculated by formula:

$$CN = \frac{NewCasesOver14DayPeriod}{Population} * 100000$$

Finally, in our study, we used the dataset, which contains the cumulative number, taken for every day and for each European country and territory in the observed period.

C. Clustering

The main task of clustering is grouping objects so that the similar objects belong to the same group (cluster). The property of a good clustering is a great similarity in a group and high diversity between groups. Similarity and difference between objects are based on data about the clustered objects.

In literature, one can find relevant information about different type of clustering (hierarchical (nested) versus partitional (un-nested), exclusive versus overlapping versus fuzzy, and complete versus partial) and different types of clusters, like Well-Separated, Prototype-Based, Graph-Based, Density-

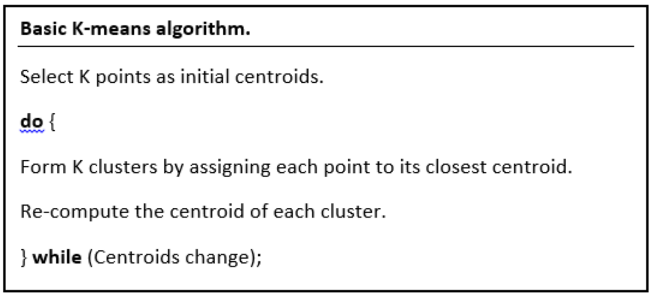
Based and Shared-Property. For more information about clustering techniques, we refer to [15].

In this research we used three different techniques for clustering:

- K-means clustering – partitional clustering;
- Agglomerative clustering, one kind of hierarchical clustering;
- BIRCH clustering – also a kind of hierarchical clustering.

D. K-means clustering

As it is mentioned, K-means is a prototype-based, partitional clustering. Partitional clustering implies dividing a set of data to disjoint subsets, i.e. every data belongs to the exactly one subset (cluster). Every cluster is a set of data which are more similar to the prototype that defines the cluster than to the prototype of any other cluster. Usually, when the data have continuous attributes, the prototype of the cluster is the mean of all points in the cluster, i.e. centroid.



Basic K-means algorithm.

```
Select K points as initial centroids.
do {
Form K clusters by assigning each point to its closest centroid.
Re-compute the centroid of each cluster.
} while (Centroids change);
```

Figure 1. Basic K-means algorithm.

In Fig. 1 the basic K-means algorithm is shown, where K is the number of required clusters. In the beginning, K initial centroids are chosen. Further, every data is associated with the most similar centroid, so every set of data gathered around the same centroid is considered as one cluster. After that, the centroid of every cluster is recomputed on the basis of associated data. This processes of association and re-computation are repeated while there exist data which change the cluster, i.e. while centroids are changing by re-computation.

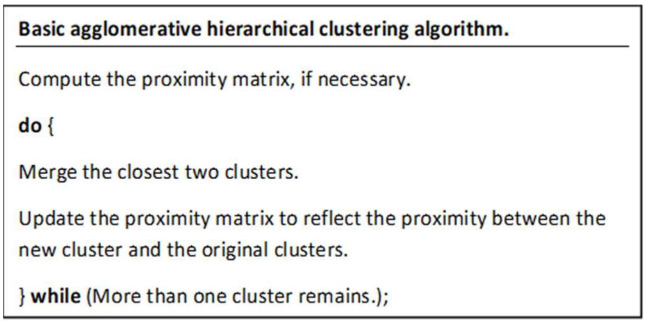
The similarity between centroid and data can be defined in various ways, depending on the goal of clustering. Also, different measures can be used. Commonly used measurements are: Euclidean (L2) or Manhattan (L1) distance for data points in Euclidean space, while cosine similarity or Jaccard measure [9] are more appropriate for documents.

More about K-means clustering can be found in [15].

E. Agglomerative clustering

Hierarchical clustering results with a set of nested clusters that are organized as a tree, i.e. it is permitted that clusters have subclusters. In such a hierarchical tree, a node represents cluster which is a union of clusters, represented by its children. Usually, the leaf nodes of the tree are singleton clusters of individual data objects.

The agglomerative hierarchical clustering, which is used in this research, forms clusters by the following principle. In the beginning, each point (every data) is an individual cluster. In each of the next steps, the closest pair of clusters are merged. The process of merging is repeated until only one cluster remains. In Fig. 2 is shown the basic agglomerative hierarchical clustering algorithm.



Basic agglomerative hierarchical clustering algorithm.

```
Compute the proximity matrix, if necessary.
do {
Merge the closest two clusters.
Update the proximity matrix to reflect the proximity between the
new cluster and the original clusters.
} while (More than one cluster remains.);
```

Figure 2. Basic agglomerative hierarchical clustering algorithm.

The cluster proximity can be defined in various ways, such as:

- MIN - the proximity between two closest points that are in different clusters, i.e. the shortest edge between two points in different clusters;
- MAX - the proximity between the farthest two points in different clusters, i.e. the longest edge between two points in different clusters;
- Group Average - the average pairwise proximities (average length of edges) of all pairs of points from different clusters.

In the case of prototype-based clustering, cluster proximity is defined as the proximity between cluster centroids. One technique to define such proximity is Ward's, which attempts to minimize the sum of the squared distances of points from their cluster centroids. This technique is used in our paper.

F. BIRCH clustering

BIRCH clustering is a scalable hierarchical clustering method which begins with generating a small and compact summary of the large dataset that retains as much information as possible [16]. This smaller piece of data is further clustered, instead of clustering the larger dataset. BIRCH clustering is based on the notation of CF (Clustering Feature) and Tree Clustering Feature (TCF). Large datasets are summarized into smaller regions, called CF entries. Each CF entry is represented by a triple (N, LS, SS) where:

- N: is the number of data points;
- LS: is the linear sum of the points;
- SS: is the squared distance of the points in the cluster.

The TCF is a very compact dataset representation. Every entry in a TCF contains a pointer to a child node and a CF entry, which is made up of the sum of CF entries in the child nodes. Each leaf node of TCF contains a sub-cluster. An overview of the BIRCH algorithm is shown in Fig. 3.

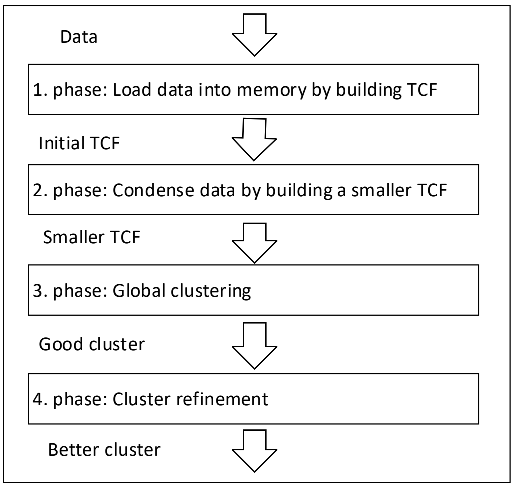

Data
1. phase: Load data into memory by building TCF
Initial TCF
2. phase: Condense data by building a smaller TCF
Smaller TCF
3. phase: Global clustering
Good cluster
4. phase: Cluster refinement
Better cluster

Figure 3. BIRCH Overview [16]

G. *Implementation of K-means, agglomerative and BIRCH clustering in Python*

In this research we used `sklearn.cluster` Python module [17].

For implementation of K-means method, `KMeans` function is used. This function has many parameters, but we will mention only three of them:

- Number of clusters;
- Starting initialization of centroids;
- Number of maximal iterations for a single run.

Default value of number of clusters is 8, but it can be user specified. In our research, we tested the performances of clustering algorithm with the number of clusters from 2 to 12.

For the starting initialization of centroids, default `k-means++` method is used. This method chooses initial centroids in order to speed up convergence. Choosing of centroids is an important step, because K-means method is sensitive to that and bad choice of initial centroids can result with poor clustering. The `k-means++` method chooses centroids distant from each other, leading to probably better results than random initialization, which is another possibility.

Centroids are chosen as follows. Let $X = \{x_1, x_2, \dots, x_n\}$ be a set of n data, $\{C_1, C_2, \dots, C_k\}$ set of k clusters and c_i centroid of the cluster C_i, for all $i \in \{1,2, \dots, k\}$.

Centroids are chosen such that minimize the inertia, or within-cluster sum-of-squares criterion:

$$inertia = \sum_{i=0}^{n} \min_{c_j \in C_j} (\|x_i - c_j\|^2)$$

One of the parameters is the number of maximal iterations for a single run. Default value of this parameter is 300, but in our research we used 1000 in order to achieve better accuracy.

In this implementation of K-means clustering, Lloyd's or Elkan's algorithm is used [17]. The average complexity of this algorithm is $O(k * n * T)$, were n is the number of samples, T is the number of iterations and k is number of clusters.

For agglomerative clustering, we used the function `AgglomerativeClustering`. We mention three parameters of this function:

- Number of clusters;
- Affinity type;
- Linkage.

Similarly as for K-means algorithm, we tested the agglomerative clustering with the number of clusters from 2 to 12. We used Euclidean affinity and Ward linkage.

BIRCH algorithm is implemented by the function `Birch`, with the following three parameters:

- Number of clusters;
- Threshold;
- Branching factor

Threshold is defined as maximum value of the radius of the subcluster obtained by merging a new sample and the closest subcluster (default value 0.5 was used). Branching factor is maximum number of CF subclusters in each node (default value equals to 50 was used).

As in previously mentioned methods, the number of clusters from 2 to 12 was used.

H. *Clustering performance evaluation*

In order to evaluate the performance of a clustering, we calculated the Silhouette Coefficient.

The Silhouette Coefficient for a single sample *i* is calculated as:

$$s(i) = \frac{b-a}{max(a,b)},$$

where a is the mean distance between the sample and all other points in the same cluster and b is the mean distance between a sample and all other points in the next nearest cluster. For a given set of samples, the value of this coefficient is the mean of the Silhouette Coefficients for each sample from that set, i.e.

$$s = \frac{1}{n}\sum_{i=1}^{n} s(i)$$

where n is the number of samples.

The smallest possible value the Silhouette Coefficient is -1 and indicates incorrect clustering. The value +1 is the largest possible and shows highly dense clustering. When clusters are dense and well separated, the value of score is higher. Values around zero indicates overlapping clusters.

IV. Experimental results

All the tests were performed on Intel(R) Xeon(R) CPU @2.30GHz and16GB RAM under the Windows 10 operating system. The version of Python compiler is 3.6.9.

Recall that we performed clustering on the dataset containing cumulative number of COVID-19 cases in the observed period. In Table 1 we show the results of clustering, for the number of clusters from 2 to 12. The table is organized as follows. The first column contains the number of clusters. The rest of the table contains the Silhouette Coefficient for K-means, agglomerative and BIRCH methods, respectively.

In Fig. 4 we graphically show the values of the Silhouette Coefficient for all three clustering algorithms and various numbers of clusters. As it can be seen, all values are higher the zero, which indicates good cluster performances.
From Table 1, one can see that the Silhouette Coefficient for all methods is not less than 0.24 and not greater than 0.39, which indicates relatively accurate clustering.

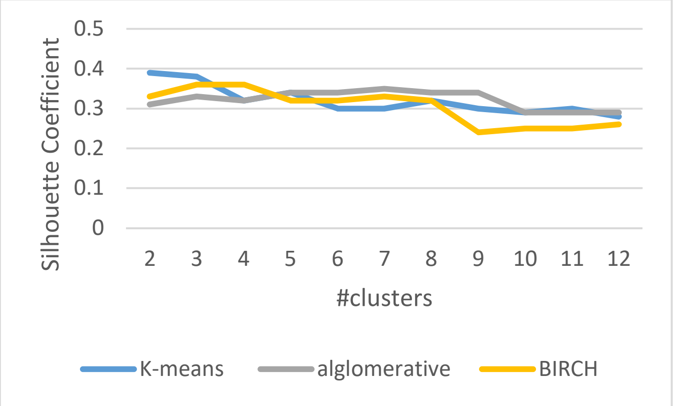


Figure 4. Graphical view of Silhouette Coefficient values

For further analysis, we decided to focus on K-means clustering and we chose one case, namely #clusters=5. It should be noted that other two clusterings obtain similar results. We graphically show the results of clustering in Fig. 5.
The results of the K-means clustering, for the case #clusters=5 are as follows.

#clusters	K-means	agglomerative	BIRCH
2	0.39	0.31	0.33
3	0.38	0.33	**0.36**
4	0.32	0.32	**0.36**
5	**0.34**	0.34	0.32
6	0.3	0.34	0.32
7	0.3	**0.35**	0.33
8	0.32	0.34	0.32
9	0.3	0.34	0.24
10	0.29	0.29	0.25
11	0.3	0.29	0.25
12	0.28	0.29	0.26

Table 1. Tabular view of Silhouette Coefficient values

Cluster 1 : ['AUT', 'HRV', 'GEO', 'LIE', 'LTU', 'LUX', 'MNE', 'SMR', 'SRB', 'SVN', 'CHE']
Cluster 2: ['ALB', 'AZE', 'BLR', 'CYP', 'DNK', 'EST', 'FRO', 'FIN', 'DEU', 'GRC', 'GGY', 'ISL', 'IRL', 'IMN', 'JEY', 'LVA', 'MLT', 'MCO', 'NOR', 'RUS', 'TUR', 'UKR']
Cluster 3: ['ARM', 'BIH', 'BGR', 'FRA', 'GIB', 'HUN', 'ITA', 'XKX', 'MDA', 'NLD', 'MKD', 'POL', 'PRT', 'ROU', 'SVK', 'ESP', 'SWE', 'GBR']
Cluster 4: ['VAT']
Cluster 5: ['AND', 'BEL', 'CZE']

The countries and territories are shown by their abbreviations. From Fig. 5 one can see that there is a certain dependency between geographical closeness and belongness to the same cluster. For example, most of the Eastern European countries belong to the same cluster (colored in red). Further, three south-eastern countries (Albania, Greece and Turkey) belong to the same (red) cluster. Large countries from Western Europe (Portugal, Spain, France, Great Britain and Italy) also belong to the same cluster, together with several Middle European countries. Switzerland, Austria, Slovenia, Croatia and Serbia are clustered in the same cluster. This can be explained by an intensive people flucutation and connection between these countries. Note that most of the countries which belong to the orange cluster are those that had been highly affected by COVID-19 in the considered time-range. Most of the countries which belong to the red cluster, are those less affected by COVID-19 in the considered time-range.

V. Conclusions

In this study we performed the clustering of European countries and territories according to the cumulative relative number of COVID-19 cases in 2020. Three clustering algorithms were used: K-means partitional clustering, agglomerative and BIRCH hierarchical clustering. We used the dataset based cumulative relative number of COVID-19 cases, taken on daily basis. For each clustering method, we examined the clustering performances according to the Silhouette

Coefficient value. The obtained results indicate a good accuracy of the examined methods on the observed dataset.

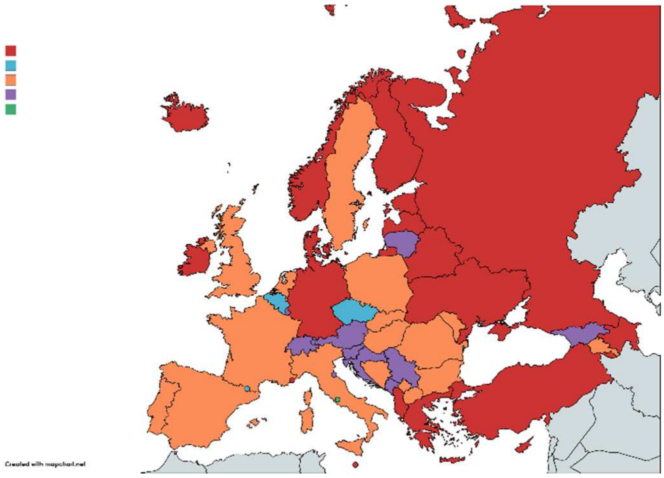


Figure 5. K-means clustering of European countries and territories with #clusters=5. Countries clustered into the same cluster are colored with the same color.

Acknowledgment

This research is partially supported by Ministry for Scientific and Technological Development, Higher Education and Information Society, Government of Republic of Srpska, B&H, under the project "Development of artificial intelligence methods for solving computational biology problems" (2020).

References

[1] N. Tayarani-and H. Mohammad "Applications of Artificial Intelligence in Battling Against Covid-19: A Literature Review." Chaos, Solitons & Fractals, (2020): 110338, in press.

[2] J. Rasheed et al. "A survey on artificial intelligence approaches in supporting frontline workers and decision makers for COVID-19 pandemic." Chaos, Solitons & Fractals, vol. 141, (2020): 110337.

[3] A. Santini "Optimising the assignment of swabs and reagent for PCR testing during a viral epidemic." *Omega*, 102341. 22 Sep. 2020, doi:10.1016/j.omega.2020.102341.

[4] L. Lampariello and S. Sagratella. "Effectively managing diagnostic tests to monitor the COVID-19 outbreak in Italy" Tech. rep. Optimization Online, 2020.url: http://www.optimization-online.org/DB_FILE/2020/03/7680.pdf.

[5] S. Mehrotra, H. Rahimian, M. Barah, F. Luo, K. Schantz "A model of supply-chain decisions for resource sharing with an application to ventilator allocation to combat COVID-19." Naval Research Logistics (NRL). 2020 May 2:10.1002/nav.21905. doi: 10.1002/nav.21905. PMCID: PMC7267382.

[6] R. Srivatsan, P. N. Indi, S. Agrahari, S. Menon and S.D. Ashok "Machine learning based prognostic model and mobile application software platform for predicting infection susceptibility of COVID-19 using healthcare data." Res. Biomed. Eng., pp. 1-12, 2020.

[7] R. M. Carrillo-Larco and M. Castillo-Cara "Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach." Wellcome Open Research, vol. 5(56), 56, 2020.

[8] L. Chaudhary and B. Singh. "Community Detection using Unsupervised machine learning technique on COVID -19 dataset." doi:10.21203/rs.3.rs-74143/v1. PPR:PPR212474.

[9] V. Zarikas, S. G. Poulopoulos, Z. Gareiou, and E. Zervas "Clustering analysis of countries using the COVID-19 cases dataset.", Data in brief, 31, 105787, (2020).

[10] N. James and M. Menzies "Cluster-based dual evolution for multivariate time series: Analyzing COVID-19. ", Chaos: An Interdisciplinary Journal of Nonlinear Science, 30(6), 061108, (2020).

[11] 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 & Fractals, 140, 110230, (2020).

[12] E. Ahlqvist et al. "Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables." The lancet Diabetes & endocrinology, vol. 6(5), pp. 361-369, 2018.

[13] S. P. Carruthers et al. "Exploring heterogeneity on the Wisconsin card sorting test in schizophrenia spectrum disorders: a cluster analytical investigation." Journal of the International Neuropsychological Society, vol. 25(7), pp. 750-760, 2019.

[14] M. Pikoula et al. "Identifying clinically important COPD sub-types using data-driven approaches in primary care population based electronic health records", BMC medical informatics and decision making, 19(1), pp. 1-14, 2019.

[15] P. Tan, M. Steinbach, A. Karpatne and V. Kumar Introduction to Data Mining (2nd Edition), Pearson, (2018).

[16] T. Zhang, R. Ramakrishnan and M. Livny "BIRCH: an efficient data clustering method for very large databases.", ACM sigmod record, 25(2), 103-114, 1996.

[17] F. Pedregosa et al. "Scikit-learn: Machine learning in Python", Journal of machine Learning research, 2011, 12: 2825-2830.