IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Analysis of the Distribution of COVID-19 in Italy Using Clustering Algorithms

Anastasiya Doroshenko
ACS Department
Lviv Polytechnic National University,
Lviv, Ukraine
anastasia.doroshenko@gmail.com

***Abstract*—This article describes the using of various clustering methods for in-depth analysis and further researches of the spread of the virus COVID-19 in Italy during February-April 2020.**

Keywords— clustering algorithms, hierarchical clustering, distance map, k-Means, data science, data mining, COVID-19.

I. INTRODUCTION

At the end of December 2019, the first cases of pneumonia of unknown origin in locals associated with the local Huanan animal and seafood market were discovered in the city of Wuhan, Hubei Province of central China. On December 31, 2019, Chinese authorities informed the World Health Organization (WHO) of an outbreak of unknown pneumonia. Since January 22, Wuhan has been quarantined; from January 24 - urban districts adjacent to Wuhan. However, despite such precautions, it was not possible to keep the virus in China. On January 31, two Chinese tourists with coronavirus who arrived in Rome were recorded in Rome. On 30 January, WHO recognized the outbreak of the new coronavirus as a public health emergency of international concern. Despite the fact that only on February 11, 2020 the disease was called the new coronavirus disease (COVID-2019), by this time the virus not only put all Wuhan residents in strict quarantine, but also spread rapidly throughout Italy. By mid-March, Italy came in second in the world in the number of cases of COVID-2019. During the pandemic, the efforts of scientists from various fields are aimed at combating COVID-2019 - medical scientists are looking for drugs, biologists are developing the genome and creating a vaccine, and specialists in the field of Data Science, using known methods and data analysis algorithms, studied the spread of the virus around the world, analyzed the data collected and used them to predict the future development of the pandemic.

II. STATEMENT OF THE PROBLEME

A. Unsufficient amount of collected historical data

To successfully use Data Mining methods, it is necessary to have a sufficient amount of collected historical data. Also, they need preprocessing - data cleaning (missing data, Noisy Data), Data Transformation (Normalization, Attribute Selection, Discretization) [2]. However, in a difficult time of pandemic, it was not possible to just wait until a sufficiently large amount of data was accumulated, because every day, the number of sick and dead from COVID-2019 is increasing. Accordingly, it was necessary to use methods that show high accuracy on small amounts of data [6].

B. Protection of patient personal data in accordance with GDPR

The second problem in studying the distribution of coronavirus was that information about patients, their age, gender, place of residence, concomitant diseases, are particularly sensitive data and cannot be disclosed according to the General Data Protection Regulation (GDPR). This problem can be solved by applying anonymization methods, etc. However, this is quite risky from a legal and ethical point of view, respectively - it is impossible to get this data quickly, in real-time, and accordingly - to get the most correct result. Accordingly, only those statistics that are public were available for analysis [1]. These are indicators such as the number of new cases per day, deaths per day, the total number of cases, the number of hospitalized people, etc.

Although the first cases were detected in the center of Italy - Rome, the epicenter of the disease was the north of the country - the province of Lombardy. Also, new cases were recorded in almost all provinces. However, in some provinces, the level of the disease remained low, while in others, after a very short time, they became “red zones” - with a high number of sick and dead. To analyze what features are common between the provinces with a high incidence rate, it was decided to cluster the provinces with the two most popular clustering methods - k-Means and hierarchical clustering.

III. COMPARISON OF CLUSTERING METHODS

Depending on the specific task, clustering methods can be used for various purposes:

- To analyze many objects and understand their structure, dividing them into groups of similar objects. As a rule, in this case, it is desirable to choose a small number of clusters.
- Reduce the amount of data stored in the case of an extra-large sample Xℓ, leaving one of the most typical representatives from each cluster. In this case, it is important not the number of clusters, but the provision of a high degree of similarity of objects within each cluster.
- Select atypical objects that do not fit into any of the clusters. This task is called a single-class classification, the discovery of atypicality or novelty (novelty detection). As a rule, in this case, individual objects that do not fit into any of the clusters are of the greatest interest.

For all these cases, hierarchical clustering can be applied - a taxonomy, when large clusters are split up into smaller ones, which in turn are split up even smaller, etc. The result of taxonomy is not only the division of many objects into clusters but also a tree-like hierarchical structure. Instead of a cluster number, an object is characterized by listing all the clusters to which it belongs, from large to small.

We will compare the effectiveness of the algorithms chosen for clustering on a sample of 1113 entries describing the epidemiological situation in Italy from February 20, 2020, to April 16, 2020, according to 17 parameters: Date, Country, Region Code, Region Name, Latitude, Longitude, Hospitalized Patients, Intensive Care Patients, Total Hospitalized Patients, Home Confinement, Current Positive Cases, New Positive Cases, Recovered, Deaths, Total Positive Cases, Tests Performed.

Data for this research was collected by Sito del Dipartimento della Protezione Civile - Emergenza Coronavirus: la risposta nazionale and uploaded into github repositories https://github.com/pcm-dpc/COVID-19 [2].

A. k-Means

The k-Means clustering algorithm is a non-hierarchical, iterative clustering method. He gained great popularity due to its simplicity, clarity of implementation and fairly high quality of work.

The main idea of the k-means algorithm is that the data is randomly divided into clusters, after which the center of mass for each cluster obtained in the previous step is iteratively recalculated, then the vectors are again divided into clusters according to which of the new centers is closer in selected metric.

The purpose of the algorithm is to separate n observations $X = \{x_1, x_2, \dots, x_n\}, x_i \in R^d, i = 1, .., n$ to k clusters $K_1, K_2, \dots K_k$, $k \in N, k \leq n$ so that each observation belongs to exactly one cluster $K_1 \cap K_2 \neq \emptyset \bigcup_{i=1}^{k} K_i = X$, located at the smallest distance from the observation $\arg min_K \sum_{i=1}^{k} \sum_{x \in K_i} \rho(x, \mu_i)^2$,

$\mu_i, i = 1, \dots, k -$ are the centers of the clusters,

$\rho(x, \mu_i)$ – is the distance function between x and μ_i.

It should be noted that the algorithm does not apply to data for which the concept of "average" is not defined, for example, categorical data.

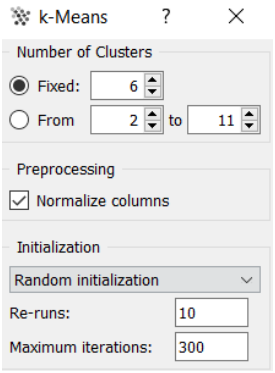

Fig. 1. Parametrs of k-Means algoritm

The advantages of the algorithm are:

- Relatively high efficiency with ease of implementation;
- High-quality clustering;
- Possibility of parallelization;
- The existence of many modifications.

The disadvantages of the algorithm are:

- The number of clusters is a parameter of the algorithm
- Sensitivity to initial conditions - initialization of cluster centers significantly affects the result of clustering.
- Sensitivity to emissions and noise - Emissions that are far from the centers of these clusters are still taken into account when calculating their centers.
- The possibility of convergence to a local optimum - an iterative approach does not guarantee convergence to an optimal solution.

For experiments, the Orange software product and installed additional widgets were used [5].

Pre-processed data is processed using the k-means algorithm with the parameters shown in Fig. 1

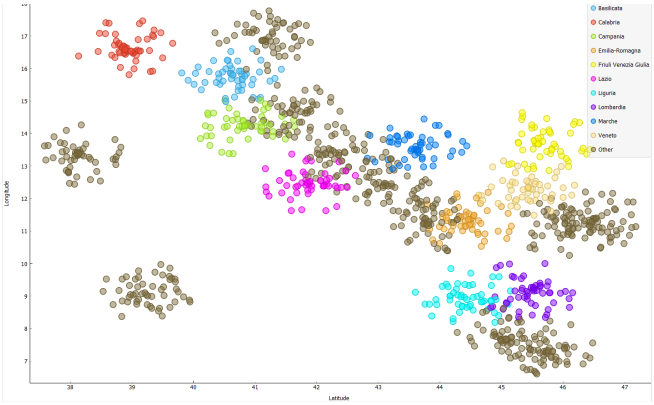

Fig. 2. The results of clustering of Italy`s regions in two-dimensional space (Latitude, Longitude).

Analyzing the obtained results, it should be noted that we obtained the best clustering results for the coordinate space (Latitude, Longitude) (fig.2).

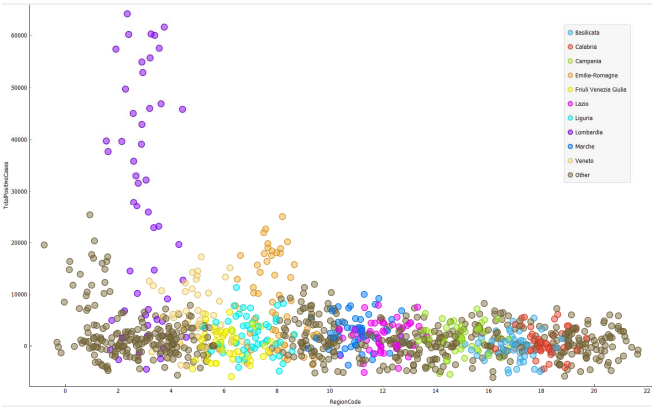

Fig. 3. The results of clustering of Italy`s regions in two-dimensional space (Region Code, Total Positive Cases).

For all other dependencies that interested us - such as the classification according to the number of sick, dead, etc., by region, only 1 cluster (Lombardy) could be distinguished, the remaining clusters overlapped and were inseparable (fig.3).

B. Hierarchical clustering

Hierarchical clustering algorithms build not one partition of the sample into disjoint classes, but a system of nested partitions. The result of taxonomy is usually represented as a taxonomic tree - a dendrogram. Among hierarchical clustering algorithms, two main types are distinguished. Divisible or top-down algorithms break down a sample into increasingly smaller clusters. Agglomerative or ascending algorithms are more common in which objects are combined into more and more large clusters.

First, each object is considered a separate cluster. For single-element clusters, the distance function $R(\{x\},\{x'\}) = \rho(x,x')$ is naturally determined.

Then the merge process starts. At each iteration, instead of the pair of the closest clusters U and V, a new cluster $W = U \cup V$. is formed. The distance from a new cluster W to any other cluster C is calculated from the distances $R(U,V)$, $R(U,C)$ and $R(V,C)$,, which by this moment should already be known: $R(U \cup V,C) = \alpha U R(U,C) + \alpha V \cdot R(V,C) + \beta R(U,V) + \gamma R(U,C) - R(V,C)$. αU, αV, β, γ - numerical parameters.

This universal equation summarizes almost all reasonable ways to determine the distance between clusters. It was proposed by Lance and Williams in 1967 [4].

In practice, the following methods are used to calculate the distances R (W, C) between the clusters W and C. For each of them, the correspondence to the Lance-Williams formula for certain combinations of parameters is proved:

- single linkage clustering or nearest neighbor distance

$$R^n(W,C) = \min_{w \in W, c \in C} \rho(w,c) \quad a_U = a_V = \frac{1}{2}, \beta = 0, \gamma = -\frac{1}{2}$$

- complete linkage clustering or Furthest neighbor distance

$$R^F(W,C) = \max_{w \in W, c \in C} \rho(w,c) \quad a_U = a_V = \frac{1}{2}, \beta = 0, \gamma = \frac{1}{2}$$

- centroid linkage clustering

$$R^C(W,C) = \frac{1}{|W||C|} \sum_{w \in W} \sum_{c \in C} \rho(w,c)$$

$$a_U = \frac{|U|}{|W|} a_V = \frac{|V|}{|W|}, \beta = \gamma = 0$$

The Orange Software also implements other algorithms for calculating the distances between clusters.

Distance Metric:

- Euclidean distance is a geometric distance in multidimensional space.
- Manhattan. According to this metric, the distance between two points is equal to the sum of the modules of the difference of their coordinates.

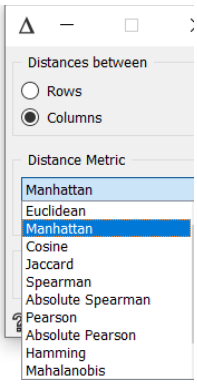

Fig. 4. The algorithms for calculating the distances between clusters in Orange.

- Cosine the similarity factor of two non-zero vectors in the pre-Hilbert space, which is calculated as the cosine of the angle between them.
- Jaccard. It's a binary measure of similarity. This method used to evaluate the similarity of finite sets, in computer science, to search for similar documents, plagiarism, etc.
- Spearman. According to this metric, the distance is linear correlation between the rank of the values, remapped as a distance in a [0, 1] interval)
- Pearson According to this metric, the distance is linear correlation between the values, remapped as a distance in a [0, 1] interval)
- Hamming. It's a binary measure of similarity. The number of features at which the corresponding values are different. [5].

The widget supports four ways of measuring distances between clusters:

- Single linkage computes the distance between the closest elements of the two clusters
- Average linkage computes the average distance between elements of the two clusters
- Weighted linkage uses the WPGMA method
- Complete linkage computes the distance between the clusters' most distant elements

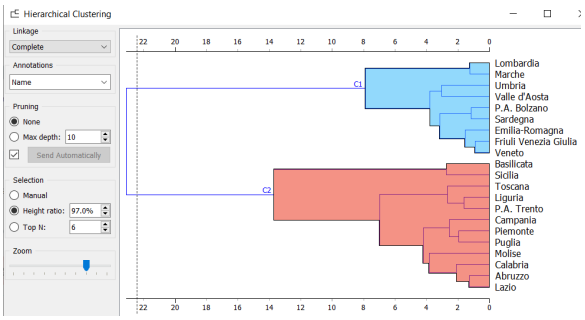

Fig. 5. Hierarchical clustering diagram with a height ratio 97 %

Huge dendrograms can be pruned in the Pruning box by selecting the maximum depth of the dendrogram. This only affects the display, not the actual clustering.

The widget offers three different selection methods:

- Manual
- Height ratio
- Top N

Applying hierarchical clustering for our test set, we obtain this results (fig.5, fig.6).

Comparing the selected clustering methods, we can note that both methods did a good job of dividing regions into clusters. However, in the k-Means algorithm, the resulting clusters describe a well-known division of regions according to their geographical location. For other criteria - similarity in the number of sick, hospitalized, dead - the accuracy of clustering is insufficient.

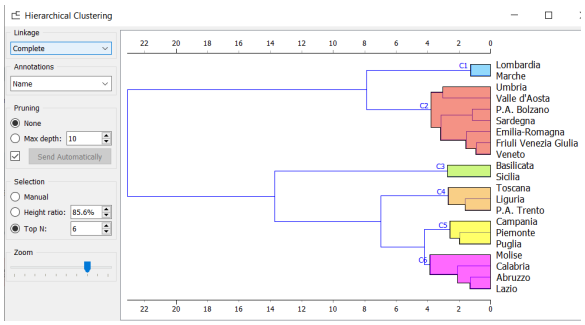

Fig. 6. Hierarchical clustering diagram for 6 clusters

At the same time, the use of hierarchical classification showed more interesting results. Given the height ratio of 97%, the two clusters formed showed the division of regions into those with high incidence (blue cluster, C1) and those with low incidence (red cluster, C2) (fig.5).

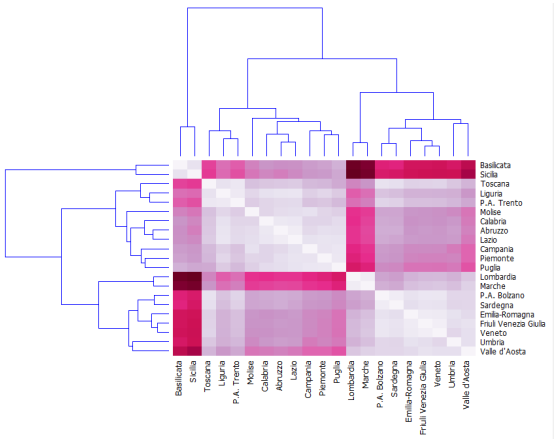

Fig. 7. Distance Map between clusters for Hierarchical clustering

Given a specified criterion — the obligatory selection of 6 clusters — we obtained clusters whose elements at first glance are not so unambiguously interconnected - for example, the number of cases in one cluster can be of the same order, but not the closest from all regions (fig.6).

CONCLUTIONS

After a detailed analysis of the resulting clusters and additional information about the industry of the Italian regions (fig.8), it was found that the clusters obtained combine regions with a similar level of industrial development and similar industries.

Thus, it became clear that the number of sick, hospitalized and deceased citizens of Italy depends on which of the industries in it is developed. Thus, the best epidemiological situation is observed in regions with poorly developed economies or in those where the key areas of activity are: agricultural, construction and trade. The worst is in regions with heavy industry and other industries.

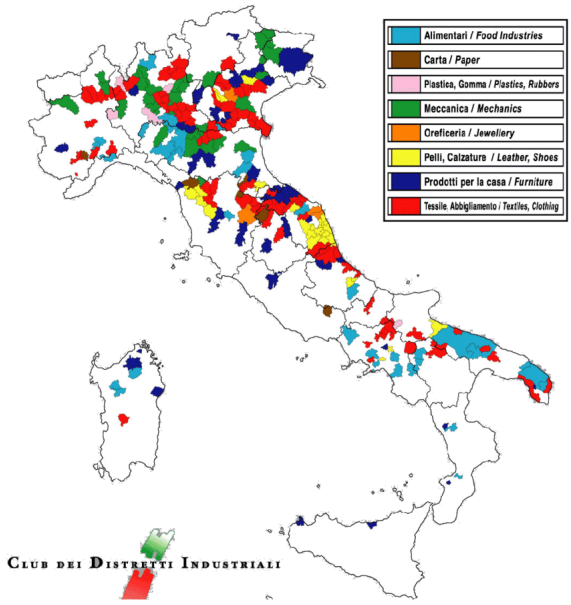

Fig. 8. Regional specialization in Italy in the macroeconomic sector

The subject of further research may be a more detailed analysis of the revealed dependence, however, it is obvious that the task of identifying new, hidden knowledge from the available data is accomplished through the use of the hierarchical classification method.

REFERENCES

[1] Sito del Dipartimento della Protezione Civile - Emergenza Coronavirus: la risposta nazionale and uploaded into github repositories https://github.com/pcm-dpc/COVID-19.

[2] O. Mishchuk, R. Tkachenko and I. Izonin, "Missing Data Imputation through SGTM Neural-like Structure for Environmental Monitoring Tasks" In: Hu, Z.B., Petoukhov, S., (eds) Advances in Computer Science for Engineering and Education II. ICCSEEA2019. Advances in Intelligent Systems and Computing. Springer, Cham, pp. 142-151, 2019.

[3] L. Rokach and O. Maimon. The data mining and knowledge discovery handbook, Springer US, 2005.

[4] G. N. Lance and W. T. Willams, "A general theory of classification sorting strategies. 1. hierarchical systems" Comp. J, no. 9, 1967, pp. 373–380.

[5] Hierarchical Clustering, Orange Widgets Catalog, Documentation https://orange.biolab.si/widgets/unsupervised/hierarchicalclustering

[6] A. Doroshenko, "Application of global optimization methods to increase the accuracy of classification in the data mining tasks", in: Luengo D., Subbotin S. (Eds.): Computer Modeling and Intelligent Systems. Proc. 2-nd Int. Conf. CMIS-2019, vol-2353: Main Conference Zaporizhzhia, Ukraine, April 15-19, 2019., pp.98-109, 2019.

[7] A. Markusen,"Sticky Places in Slippery Space: A Typology of Industrial Districts," Economic Geography, Taylor & Francis Journals, vol. 72(3), pages 293-313, July 1996.

B. Barone and C. Brasili, "La specializzazione e l'internazionalizzazione dell'industria manifatturiera delle regioni italiane". In book: Rapporto ICE 2010-2011 L'Italia nell'economia internazionale. Chapter: 7 - Il territorio Publisher: Istitito Commercio Estero (ICE), Ministero dello Sviluppo Economico, 2011.