Unsupervised machine learning and pandemics spread: the case of COVID-19

Roberto F. Silva[1], Fernando Xavier[1], Antônio M. Saraiva[1], Carlos E. Cugnasca[1]

[1]Departamento de Engenharia de Computação e Sistemas Digitais – Escola Politécnica
Universidade de São Paulo (USP) – São Paulo – SP – Brazil

{roberto.fray.silva,fernando.xavier}@gmail.com, {saraiva, carlos.cugnasca}@usp.br

Abstract. *Epidemics have severe impacts on people's health. The COVID-19 has infected more than 3 million people in 3 months. In this work, we explore the use of unsupervised machine learning to evaluate and monitor the disease spread worldwide in three points in time: January, February, and March of 2020. Besides the features related to the disease spread, we consider HDI, population density, and age structure. We define the number of clusters using the elbow and agglomerative clustering methods, then implement and evaluate the k-means algorithm with 3, 4, and 5 clusters. We conclude that four clusters better represent the data, analyze the clusters over time, and discuss the impacts on each depending on the measures adopted.*

1. Introduction

An epidemic represents a situation in which a disease affects a large number of people in a short period, generating a peak in demand for health services to stop the disease spread and treat the sick people. In 2019, Brazil went through a severe measles epidemic, with 11,000 cases reported in the State of Amazonas [Cabral et al., 2019].

On March 11, the World Health Organization (WHO) declared the disease caused by this new coronavirus to be a pandemic, in which the disease was present in 114 countries, with 118,000 cases and 4291 deaths worldwide [WHO, 2020; Dong, Du, and Gardner, 2020]. According to the WHO, there are more than 3 million cases and more than 200,000 deaths, with 210 countries and territories affected by COVID-19 [Dong, Du, and Gardner, 2020].

The research is divided into: (1) understanding the disease; (2) developing forms of treatment; (3) discovering ways to decrease the velocity of the spread; and (4) developing immunization methods. The same happened with other diseases, such as Severe Acute Respiratory Syndrome (SARS), in 2002-2003, and H1N1, in 2009-2010.

The spread of infectious and contagious diseases is a function of several variables: the environment, socioeconomic conditions, mobility, among others. AI techniques have been applied to analyze this data and extract useful information. In the 2009 pandemic caused by H1N1, the work by Attaluri et al. (2009) used machine learning techniques to classify H1N1 viral strains.

Clustering encompasses a group of techniques that are part of unsupervised machine learning. Its main objective is to identify the groups that better separate the data points in a given dataset, based on the distance between those points in n-dimensional space [Ghahramani, 2003]. The k-means is the most used clustering technique on a diverse set of domains [Jain, 2010; Steinley, 2006].

According to [Jain, 2010; Steinley, 2006], this technique has three main steps: (i) create points that are used as cluster centers on n-dimensional space; (ii) associate all points in the dataset with the closest clusters; and (iii) recalculate the cluster centers and repeat step (ii).

There are several interesting applications of k-means in the healthcare domain. The work by [Haraty, Dimishkieh, and Masud, 2015] proposed an extension of the algorithm focusing on healthcare data. [Martis et al., 2014] used k-means and genetic algorithms to identify patients with arrhythmia. The research by [Stricker et al., 2013] had satisfactory results on using principal component analysis and k-means to identify dietary patterns and predict coronary heart disease and stroke.

To the best of our knowledge, this is one of the first research pieces to evaluate the use of clustering techniques to try to find insights into how the disease evolved in different countries. Knowledge about the spreading of a disease is essential for developing countries with fewer resources, both in terms of health services and infrastructure.

The main objective of this work was to use clustering, an unsupervised learning technique, to identify and evaluate the ideal number of clusters of countries regarding COVID-19. We considered three periods in time, which reflected the disease spreading worldwide, as well as features that are specific for each country, such as HDI, population density, and age structure. We also evaluated world maps based on the clustering results. Lastly, we identified the leading countries in each cluster and evaluated how they might change over time, depending on the measures adopted.

This work is organized as follows: Section 2 describes the methodology adopted; Section 3 contains the main results and discussions on these results; and Section 4 contains the concluding remarks.

2. Methodology

The methodology used in this work was composed of five steps:

1. Data gathering from official databases, for three periods: 01/23/2020 (initial stage, 6 countries), 02/19/2020 (increasing importance, 26 countries), and 03/16/2020 (pandemics, 127 countries). The data sources were: World Bank[1]: total population, population density and age structure in 2018; United Nations Development Programme[2]: HDI in 2018; World Health Organization COVID-19 Situation Reports[3], collected from Our World in Data[4]: total number of cases, the total number of deaths, new cases, and new deaths; and the Johns Hopkins University CSSE [Dong, Du, Gardner, 2019][5]: number of recovered patients. We collected the data on 03/16/2020;

2. Data preprocessing and analysis. We preprocessed the three datasets (one for each period) to remove missing data. Then, we calculated the health indicators. We analyzed all features in terms of frequency distribution, presence of outliers, and their overall pattern on the three datasets. Lastly, we have developed the final datasets. These consisted of ten features for each country: total cases; age structure with three categories: less than 14, from 15 to 64, and more than 65 years old; prevalence;

1 http://wdi.worldbank.org/table/2.1#
2 http://hdr.undp.org/en/data#
3 https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports
4 https://ourworldindata.org/coronavirus-source-data
5 https://github.com/CSSEGISandData/COVID-19

incidence; the average incidence in the last seven days; death rate; density; and HDI. The health indicators were calculated per 100.000 people. As the features were in different ranges, we used the standard scaling method to standardize each feature;

3. Implementation of agglomerative hierarchical clustering and the elbow method for determining the optimal number of clusters, considering 1 to 30 clusters;

4. Implementation and analysis of the k-means++ for each dataset. Based on the results of Step 3, we identified the best options to be 3, 4, and 5 clusters. We then implemented these three options for each dataset. We conducted several experiments varying the input variables and the model's hyperparameters and implemented PCA-2 and PCA-3 to better evaluate the features and illustrate the separation of the clusters.

5. Model results analysis, considering an analysis of each cluster's compositions and an analysis of the world maps generated for each dataset. We also analyzed how the actions of the leading countries in each cluster group affected their position changes from one period to the next.

We used the following Python libraries: Pandas, Scikit-Learn, Matplotlib, NumPy, Seaborn, SciPy, and Geopandas. The implementation was done using a Google Colab TPU. The datasets and the code are available on an open Github repository[6].

3. Results and discussion

Figure 1 illustrates the dendrogram and the elbow method implementations for the March dataset. Both methods point out that the number of clusters is between 3 and 5. We observed the same pattern when we implemented the PCA-2 and PCA-3 transformations on the datasets.

Considering all experiments, we can infer that 4 is the most likely optimal number of clusters. We observed this result for both February and March datasets. Nevertheless, in January, as there were few data available, no definite conclusions could be reached as to the optimal number of clusters for that period. The January dataset contained only 581 cases and five countries. China corresponded to 574 cases.

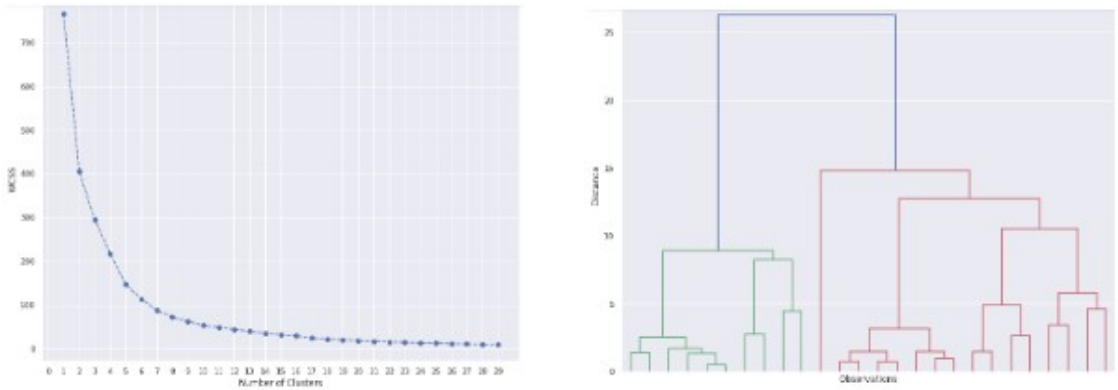

Figure 1. Elbow method (left) and dendrogram (right) for the March dataset.

As the February and March datasets had more data and presented better clustering results, we will focus on their analysis and discussion. Figure 2 illustrates, on the left side, the four clusters identified by the algorithm for February. These are: China

6 https://github.com/rfsilva1/covid19clustering

(orange); Singapore (green); South Asia and North of Africa (red); and North America, Europe, Russia, Japan, South Korea, and Australia (blue).

Figure 2. World maps illustrating the four clusters: (i) on the left, for the February dataset; and (ii) on the right, for the March dataset.

China is the only member of the first cluster, as it was the country that suffered the most at the beginning of the disease. This happened for two main reasons: the newness of the disease, and its fast spread, as mentioned in section 2. To avoid worsening the situation, the immediate measures that were adopted were: (i) social distancing and quarantine; (ii) construction of hospitals and redesign of the health system; (iii) isolation of the Wuhan province.

Singapore is the only member of the second cluster because its population density is the highest in the dataset. This result can have severe impacts if the disease spreads as in Wuhan. To avoid this from happening, the government: (i) tracked individuals that might have been exposed; (ii) tested the population; (iii) suspended events and gatherings; and (iv) and enforced quarantines and social distancing.

The third cluster contained countries that started to present the first impacts of the disease, but that was still far from having an impact as high as the first cluster. Nevertheless, their actions would define if they would move towards the first or the last cluster, which was composed of the countries that still did not show significant impacts in terms of incidence, prevalence, and the number of deaths.

Figure 2 illustrates, on the right side, the four clusters identified by the algorithm for March. These are: Italy, Estonia, Iceland, and Switzerland (red); Singapore (green); Iran, North, Central and Western Europe (orange); and the rest of the world (blue).

The first cluster represents countries that are in the most challenging situations, with a higher number of total cases and deaths in relation to their populations. These countries must enforce population quarantine and isolation of sick and older people to reduce the disease’s impact. Nevertheless, as in China's case, they may still face weeks of an increase in incidence, prevalence, and the number of deaths.

The second cluster, as in the February dataset, contains only Singapore due to its high population density. Both clusters can have devastating impacts on the disease soon. The first, due to its widespread and the lack of preemptive measures. The second, due to the possible impact of spread due to its high population density. This cluster should continue to enforce population quarantine and testing.

The third cluster represents countries with a high number of cases and deaths in which the disease is not as widespread as the countries in the first cluster. Nevertheless, one can infer, based on results from the February and March clusterings, that these

countries are moving towards the first cluster, unless they enforce measures to reduce its spread, such as social distancing for its population and isolation of the older people.

The last cluster contains both countries in which the disease is starting its impacts (such as in North and South America, Asia, and Oceania) and countries that are currently stabilizing or reducing the impacts of the disease (such as China, Japan, and South Korea). The countries in this cluster should increase population testing for the disease and adopt a quarantine for at least two weeks to better evaluate their prevalence and incidence.

The cases of China, South Korea, and Japan are particularly interesting. The first had the most devastating effects of the disease on the dataset. Nevertheless, it moved to this cluster due to the reduction in its prevalence and incidence. The other two have taken measures that seem to be enough to reduce the immediate impact of the disease, reduce incidence and prevalence, and contain its spread.

The clustering techniques can be used, together with other simulation models, to help predict the impact of disease in a specific country or cluster. It can also be used to identify which countries made decisions that significantly reduced the disease's impact. At the moment, these are mainly: China, Japan, South Korea, and Singapore.

The main limitations of this work are: (i) the lack of data available, especially in the beginning of the spread; (ii) the lack of features that point out to a country recovery; and (iii) the lack of a method to estimate if some countries may have more cases than reported, due to lack of testing.

4. Conclusions

Pandemics such as COVID-19 have a severe impact on people's lives. In 3 months, this disease has already infected about 3 million people, with more than 200.000 deaths. As the disease dynamics are still unknown, especially in the environments in developing countries, it is crucial to estimate the possible impacts of the disease. The disease evolution should be monitored continuously to allow for better decision-making.

In this work, we used the k-means algorithm to divide the countries into clusters regarding the spread of COVID-19 in three periods, considering several features. We concluded that four clusters better describe the datasets. We identified the countries in each cluster, illustrated and analyzed the data on worldwide maps, and made predictions on how they may evolve, depending on the adopted measures. We believe that the methodology could be applied every day, in an online mode, to evaluate the disease's evolution in different countries. This is especially important for developing countries to gain insights that can help with better decision-making.

Further works should: (i) incorporate more data, as it is released; (ii) evaluate the use of data gathered from other pandemics such as SARS; (iii) incorporate more socio-economic data features, such as population mobility; (iv) consider other unsupervised learning techniques, such as self-organizing maps; and (v) with more data available, start to use supervised learning techniques that incorporate time series, such as long short-term memory networks to predict the impacts of the disease better.

Acknowledgements

This work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001, Itaú Unibanco S.A. through the Itaú

Scholarship Program, at the Centro de Ciência de Dados (C2D), Universidade de São Paulo, Brazil, and also by the National Council for Scientific and Technological Development (CNPq).

References

Attaluri, P. K., Zheng, X., Chen, Z., Lu, G. (2009) "Applying machine learning techniques to classify H1N1 viral strains occurring in 2009 flu pandemic". BIOT-2009, v.21.

Cabral, M. C. et al. (2019) "Epidemia de sarampo e vacinação de bloqueio: um diagnóstico situacional dos estados do Amazonas, Roraima e Pará". Revista Saúde e Meio Ambiente, v.9, n.3, p. 1-7.

Dong, E., Du, H., Gardner, L. (2020) "An interactive web-based dashboard to track COVID-19 in real time". *The Lancet Infectious Diseases, Correspondence,* p.1-2.

Ghahramani, Z. (2003) "Unsupervised learning". In: Summer School on Machine Learning, p. 72-112. Springer, Berlin, Heidelberg.

Haraty, R.A., Dimishkieh, M., Masud, M. (2015) "An enhanced k-means clustering algorithm for pattern discovery in healthcare data". International Journal of distributed sensor networks, v.11, n.6, p.615740.

Jain, A.K. (2010) "Data clustering: 50 years beyond k-means", Pattern Recognition Letters, v.31, n.8, p.651-666.

Martis, R.J., Prasad, H., Chakraborty, C., Ray, A.K. (2014) "The application of genetic algorithm for unsupervised classification of ECG". In Machine Learning in Healthcare Informatics, p. 65-80. Springer, Berlin, Heidelberg.

Steinley, D. (2006) "K-means clustering: a half-century synthesis". British Journal of Mathematical and Statistical Psychology, v.59, n.1, p.1-34.

Stricker, M.D. et al. (2013) "Dietary patterns derived from principal component-and k-means cluster analysis: long-term association with coronary heart disease and stroke". Nutrition, Metabolism and Cardiovascular Diseases, v.23, n.3, p.250-256.

WHO. WHO Director-General's opening remarks at the media briefing on COVID-19, WHO, 2020. Available in: <https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020>. Accessed in: March 15, 2020.