2020 IEEE 6th International Conference on Control Science and Systems Engineering

A Hybrid Clustering Model for Analyzing COVID-19 National Prevention and Control Strategy

Yong Shuai, Chunxu Jiang, Xinyi Su, Can Yuan, Xiaoping Huang
Chongqing Ceprei Industrial Technology Research Institute
Chongqing Key Laboratory of Reliability Technology for Smart Electronics
Chongqing, China
e-mail: alexshuai@sina.com, 553755652@qq.com, vickysxy019@gmail.com, 766543078@qq.com, 2250584349@qq.com

***Abstract*—In order to analyze the relationship between the national epidemic strategy response to COVID-19 and the severity of the current epidemic, this paper firstly collects the national strategic data responds to the COVID-19 epidemic situation and the resulting data of COVID-19 in these nations, then use StandardScaler normalization method to pre-process the data, and use K-Means Clustering, Agglomerative Clustering and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering algorithm to perform cluster analysis on the above two types of data, use Silhouette Coefficient index , Calinski Harabasz Score, Davies bouldin score to evaluate the clustering results, and take the clustering results with the best clustering effect to analyze the impact of national epidemic policies on the current epidemic results, finally use cases to prove the effectiveness and feasibility of the model, and provide data support and strategic support for the prevention and control of the epidemic similar to COVID-19 in the future.**

Keywords-tCOVID-19; hybrid clustering model; no label clustering performance index; prevention and control strategy

I. INTRODUCTION

Since 2020, countries around the world have been attacked by new coronaviruses. Compared with SARS (atypical pneumonia) in 2003, the outbreak of the epidemic showed a stronger spread and the number of infected people increased sharply, which posed a great challenge to disease prevention and control in various countries. As of May 23, 2020, there have been more than 5 million confirmed cases worldwide. Governments of various countries have adopted various positive or negative policies (such as avoiding assembly, closing ports, and strict home management measures, etc) to deal with the epidemic due to their judgment to the epidemic and their own ruling needs, and these policies have important impact on the outcome of the epidemic.

To this end, this paper collected and pre-processed some national strategy data of COVID-19 epidemic situation and the current national COVID-19 infection, death and other results data, use multiple clustering algorithms to perform cluster analysis on the above two types of data, and use No label Clustering performance index to evaluate the clustering results, and finally use cases to analyze the impact of national epidemic policies on the current epidemic results and support countries to prevent and control the epidemics similar to COVID-19 in the future.

II. MODELING IDEAS

A. Overall Modeling Ideas

The overall modeling ideas of this paper include data collection, data preprocessing, data clustering, clustering result evaluation and clustering result association analysis, as shown in the following figure.

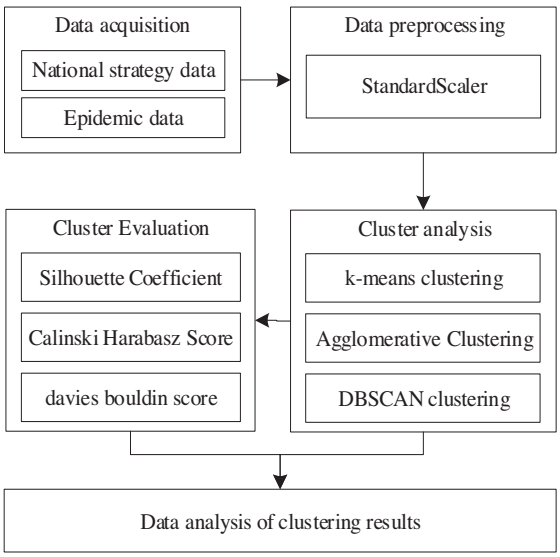

Figure 1. Overall modeling ideas.

B. Data Collection

The data collection includes national strategy data and result data in response to the COVID-19 epidemic.

1.National strategy data

The national strategy data attributes and descriptions are shown in Table I.

TABLE I. NATIONAL STRATEGY EXPLANATION

No	Attributes	Explanation
1	Close border crossing	Refers to the closure of all border crossing
2	Restricted flight	Restrict flights on international and domestic flights
3	Close public gathering place	Such as closing movie theaters, restaurants, and other gathering places
4	Limit large gatherings	Limit a large number of people to gather
5	Strict home management measures	Except for people who must go out, nationals are prohibited from going out
6	Mass immunization policy	The percentage of animals in a group that are resistant to pathogen infection reaches a certain percentage, so that the entire group can be resistant to the pathogen
7	Wear a mask in public	Government forces nationals to wear masks in certain public places
8	Increase social distance	The government forces citizens to increase social distance in certain public places
9	Feudal city	Only allowed to enter the city, not allowed to leave the city, bus, subway, ferry are temporarily suspended to avoid collision of people.
10	Increase the number of tests	Based on the existing COVID-19 detection capability, the detection capability is greatly improved

2.Linked data of epidemic results

The epidemic result related data and its calculation method are as follows

(1)The cure rate, which refers to the average number of COVID-19 per 100 patients that can be cured, reflecting the probability of being curable. The calculation method is: cure rate = total number of cured COVID-19 / total number of infected COVID-19 * 100%.

(2)Mortality, which refers to the average number of deaths per 100 patients of COVID-19. The calculation method is: mortality = total number of COVID-19 deaths due to infection / total number of COVID-19 infections * 100%.

(3)Health Accessibility Quality(HAQ) derived from literature [3], which is a comprehensive indicator to evaluate a country's medical level, including the development level of medical level, medical policy, medical quality, and the balance of medical development.

(4)The degree to which the case curve is flattened is an important factor in judging a country's fight against the epidemic. In response to the COVID-19 epidemic, in the absence of special drugs and vaccines, the best way to treat patients is to concentrate the best medical resources, let the patients survive, and overcome the virus through the patients' own immunity. Therefore, when the number of virus infections exceeds the upper limit of medical resources, many people will die because they have not received timely treatment. At the same time, the growth rate can be slowed by reducing the average number of cases produced by a single case, and it is possible to treat as many COVID-19 patients as possible within the scope of medical resources and reduce mortality.

For the measurement of the level of COVID-19 flattening, this article first divides it into five stages, namely the rapid rise period, the gentle rise period, the top volatility period, the rapid decline period and the gentle decline period, as it is shown in Figure 2.

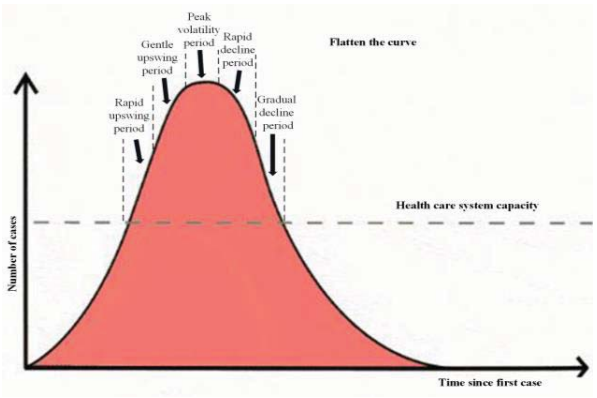

Figure 2. Case flattening curve.

The value ranges of each stage are [0, 0.2), [0.2, 0.4), [0.4, 0.6), [0.6, 0.8), [0.8-1]. The basic situation of each stage and the corresponding value range are shown in TABLE II . Among them, P1 and P2, P4 and P5 may appear alternately.

TABLE II. CASE FLATTENING CURVE RANGES

No	period	Overview	Ranges
P1	Rapid rise	The COVID-19 epidemic broke out. Due to untimely government measures and insufficient public awareness, the number of daily infections has increased rapidly.	[0,0.2)
P2	Gentle rise	Due to the influence of the government's positive measures and the enhancement of people's awareness, the rise in the number of daily infections has slowed	[0.2,0.4)
P3	Peak volatility	The number of daily infections reached the peak, with small fluctuations	[0.4,0.6)
P4	Rapid decline	Effective government measures and active actions of the public have reduced the infection rate, and the number of daily infections has begun to decline rapidly	[0.6,0.8)
P5	Gradual decline	The decline in the base of infection, coupled with government measures and the active actions of the people, the number of daily infections has slowly declined until it is completely controlled	[0.8-1]

C. *Data Preprocessing*

The principle of standardization using StandardScaler is to subtract the mean of the data by its attributes (by column) and then divide by its variance. The final result is that for each attribute / column all data are clustered around 0, and the variance value is 1. The calculation method is $x^* = \frac{x - \mu}{\sigma}$, Where μ is the average of all sample data, σ is the standard deviation of all sample data.

D. Cluster Analysis

The clustering algorithm used in this article includes K-mean Clustering, Agglomerative Clustering and DBSCAN Clustering. Because the above three algorithms are typical clustering algorithms, so this article only lists the basic idea of each algorithm. As it is shown in the TABLE III.

TABLE III. CLUSTERING ALGORITHM OVERVIEW

No	Algorithm	Overview
1	K-mean Clustering	The K-Means clustering algorithm is a simple iterative clustering algorithm that uses distance as a similarity measure. Therefore, the given data set is divided into K classes, the center of each class is obtained by the average of all points in the class, and each class is represented by the cluster center.
2	Agglomerative Clustering	Hierarchical clustering is divided into two categories according to the way of clustering: top-down and bottom-up. The bottom-up clustering algorithm initially assumes that each sample is a separate category, and then successively merges the categories until there is only one category at the end. In the end, you will get a structure similar to a tree. The root of the tree is a category, which contains all the sample points, and the leaves are a cluster of only one sample.
3	DBSCAN Clustering	DBSCAN is a density-based clustering algorithm, which assumes that the types of samples can be distinguished by the closeness of the sample distribution. Samples of the same category are closely connected; samples of different categories are distributed relatively far away. By classifying closely connected samples into one category, a clustering category is obtained. By dividing all the closely connected samples into different categories, we get the final results of all clustering categories.

E. Cluster Evaluation

The cluster evaluation index uses Silhouette Coefficient, Calinski Harabasz Score, Davies bouldin score. The calculation formula of each cluster evaluation algorithm are shown in TABLE IV.

TABLE IV. CLUSTER EVALUATION INDEX OVERVIEW

No	Cluster evaluation index	Calculation formula
1	Silhouette Coefficient Index	$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$
2	Calinski Harabasz Score	$S(k)=\frac{Tr(B_k)}{Tr(W_k)}\times\frac{N-k}{k-1}$
3	Davies bouldin score	$DBI=\frac{1}{N}\sum_{i=1}^{N}\max_{j\neq i}\left(\frac{\bar{S}_i+\bar{S}_j}{\|w_i-w_j\|_2}\right)$

III. CASE ANALYSIS

This project collected the strategy data and results data of the COVID-19 epidemic in 27 countries till May 13, 2020. In order to avoid the author's subjective judgment, this article refers to these countries as N0-N26. Use StandardScaler normalization method to preprocess the data, and use K-Means clustering, agglomerative clustering and DBSCAN clustering algorithm to perform cluster analysis on the above two types of data, the evaluation values are shown in TABLE V .

TABLE V. CLUSTERING EVALUATION RESULT

Cluster evaluation index	Clustering results of epidemic result data		
	DBSCAN Clustering	***K-mean Clustering***	***Agglomerative Clustering***
Silhouette Coefficient	-0.092944468	-0.071251685	-0.076316786
Calinski Harabasz Score	0.537236491	1.272349272	1.122302158
Davies Bouldin Index(DBI)	5.450491722	3.699639413	4.406963231
Cluster evaluation index	**Clustering results of national strategy data**		
	DBSCAN Clustering	***K-mean Clustering***	***Agglomerative Clustering***
Silhouette Coefficient	0.221374353	0.297923365	0.251924821
Calinski Harabasz Score	4.549846887	9.043263288	7.189368771
Davies Bouldin Index(DBI)	1.81229113	1.277461624	1.278812197

It could be seen from the above table that the optimal algorithm for clustering results is K-Mean clustering. The distribution graph of K-Mean clustering clustering results are shown in Figure 3 and Figure 4.

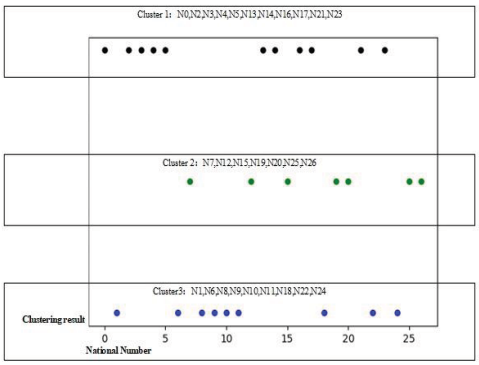

Figure 3. K-Means clustering graph of infection result data.

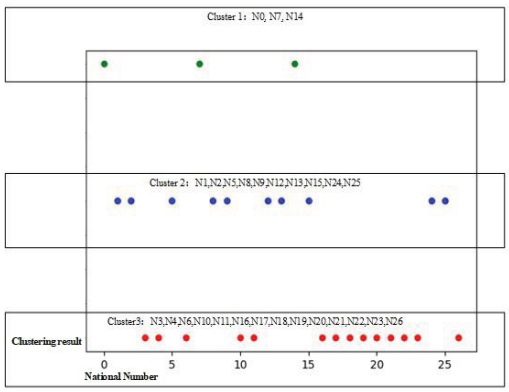

Figure 4. K-Means clustering graph of National strategy data.

Therefore, we analyze the clustering results of K-Mean clustering and find out the conclusions as below.

(1) From Figure3 ,we can see that since the N7 and N12 countries have taken active prevention and control measures, made every effort to improve the virus detection level of the country, and have good HAQ values, they can quickly smooth the case curve and achieve a good effect in fighting against the epidemic.

(2)From Figure 3 ,we can see that although countries N4 and N114 have adopted group immunization measures, the effect is not great, and there is no obvious difference from countries that have not taken such measures

(3)From Figure4 ,we can see that N13 and N25 are in the same category. Although both countries have adopted strict control measures, due to their low medical level, the control effect of the epidemic situation is not ideal.

(4)From Figure 3 and Figure 4 ,we can see that although N0, N3, N4, N5, N6, N14, N15 belong to developed countries in Europe and America, and their medical level is similar to the basic level of national quality, because different countries have adopted different strategies, their epidemic prevention and control have also produced very different results.

At the same time, we can find that the main national strategic attributes affecting the flattening curve of mortality and epidemic cases in various countries are strict home management measures and increasing the number of tests.

IV. Conclusion

In this paper, a variety of clustering algorithms and label-free clustering evaluation indicators are used to establish a hybrid clustering model. The analysis of COVID-19 epidemic strategy data and result data will play a positive role in the future prevention and control of COVID-19 epidemic. At the same time, because the national strategy data adopted in this paper is the data up to a certain point in time, but in the actual process, the early implementation of some strategies that are beneficial to epidemic prevention and control will seriously affect the mortality and the case flattening curve. These need to be improved in the future modeling.

Acknowledgment

The first two authors contributed equally to this paper. This work was supported by the Project of Integrated Standardization and New Model Application of Intelligent Manufacturing by Ministry of Industry and Information Technology (MGY1804080), Chongqing Science and Technology Bureau Project (cstc2019jscx-fxyd0298).

References

[1] Imperial College COVID-19 Response Team, “Impact of Non-pharmaceutical Interventions (NPIs) to Reduce COVID-19 Mortality and Healthcare Demand“, pp.1–20, 16 March 2020.

[2] S. Eubank, I. Eckstrand, B. Lewis, S. Venkatramanan, M. Marathe, C. L. Barrett. “Commentary on Ferguson, et al., ‘Impact of Non-pharmaceutical Interventions (NPIs) to Reduce COVID-19 Mortality and Healthcare Demand’”. Bulletin of Mathematical Biology, Vol.82,pp.1-7, April 2020.

[3] GBD 2016 Healthcare Access and Quality Collaborators, “Measuring performance on the Healthcare Access and Quality Index for 195 countries and territories and selected subnational locations: a systematic analysis from the Global Burden of Disease Study 2016”. THE LANCET, Vol.391, pp.2236-2271, 2018

[4] Zhang Y.L, Zhou Y.J, “Review of clustering algorithms”. Journal of Computer Applications, Vol.39, pp.1869－1882 July 2019

[5] MARCIN KOZAK, “'A Dendrite Method for Cluster Analysis' by Caliriski and Harabasz: A Classical Work that is Far Too Often Incorrectly Cited”. Communications in statistics: theory and methods, Vol.3pp. 2279-2280, November 2012

[6] Longford Nicholas T,Bartosova Jitka, “A confusion index for measuring separation and clustering”. Statistical Modelling, Vol.14, pp.229-255, March 2014