A 3M K-means algorithm for fast and practicably identifying COVID-19 close contacts

Xun Liang[1], *Zihuan* Feng[1,*], *Yang* Xue[1], and *Xiaolei* Zhao[1]

[1]School of Information, Renmin University of China, Beijing, 100872, China

Abstract. Given that the risk of the COVID-19 epidemic still exists and the flow of patients is difficult to monitor, identifying the people who have had close contact with the confirmed cases is important in anti-epidemic tasks whether in areas where the epidemic is developing rapidly or in areas where the epidemic has been phase-controlled. This article discusses how to locate people who have been in close contact with confirmed cases quickly and determine the risk of infection. From the perspective of the government, this work proposes a multi-snapshot multi-stage minority K-means (3M K-means) algorithm. This algorithm reduces the amount of data and considerably improves the speed of clustering by quickly ignoring the excluded risk classes and points in the process in the early stages, whereas traditional algorithms involve with O(N2) computational complexity which needs several days, impracticably for the COVID-19 urgent situations. The 3M algorithm greatly cuts down the computational time, thereof making the rapid warning of close contacts practicable. The methods are simple, yet efficient and practicable for the COVID-19 urgent situations The use of this algorithm can help control the COVID-19 epidemic, achieve significant cost savings, and provide the psychological guarantee of people for work resumption.

1 Introduction

The coronavirus disease 2019 (COVID-19) epidemic has affected more than 200 countries and regions in the world [1], and the situations in several countries have gradually passed the outbreak period. Although the most severe period has passed, effective control of the epidemic is a crucial step in preventing the disease from rebounding. Given the extremely strong infectivity of syndrome – coronavirus 2 (SARS-CoV-2), this disease may persist and spread in cities and densely populated urban areas for a long time until a vaccine is created. Each country's effort to control the domestic situation protects citizens and contributes to global anti-epidemic actions.

For the first time since January, Chinese health officials reported no new deaths from COVID-19 on April 6, 2020 [2]. The number of newly confirmed cases and deaths per day in China has considerably decreased, and most of the newly confirmed cases are from abroad. Provinces and cities in China have implemented different prevention and control measures in accordance with their respective situations. For example, many provinces and

* Corresponding author: fengzihuan@ruc.edu.cn.

cities have implemented the "health QR code" policy [3] and used the health QR code as an electronic voucher for allowing individuals to move around the local area. This voucher must be presented when entering or leaving a community, bus, office building, and other areas. The government can monitor the movement of citizens with the health QR code. If people have been to areas with a serious epidemic situation, the authorities will integrate big data information according to the specific situation and mark different colors on these people's health QR codes to remind them that they need to be isolated. For areas where the epidemic situation is serious, the regional government has organized house-to-house investigations and other actions to determine if asymptomatic infections exist to prevent the epidemic from worsening. Moreover, local governments obtain the location information of each user through mobile communication operators to monitor the urban internal flow of people. A user that flows across provinces or regions is reminded to self-isolate. With the support of current big data, governments at all levels have implemented various policies based on big data. Governments obtain data and technical support for epidemic prevention and control by dividing epidemic risk areas and obtaining people's location, mobility, contact persons, and other relevant information.

Researchers worldwide have conducted corresponding research on COVID-19 [4-6]. Current research focuses on the prediction of the development of the epidemic situation [7] aside from medical issues [8,9]. Cho also pointed out that the use of artificial intelligence system can help sniff out coronavirus outbreaks [10] AI will play an important role in epidemic prevention and control.

SARS-CoV-2 is highly infectious, can spread through aerosols and infect people, and has a long incubation period. Therefore, the epidemic is extremely likely to spread. A patient who does not know that he or she carries the virus and has close contact with others before diagnosis and an asymptomatic infected person who has traveled on multiple vehicles in a short time can infect people at an exponential rate, and the resulting situation will be difficult to control. Therefore, given that the risk of the COVID-19 epidemic still exists and the flow of patients is difficult to control, identifying the people who have had close contact with confirmed cases is important for controlling the epidemic whether in areas where the epidemic is developing rapidly or in areas where the epidemic has been phase-controlled.

2 Algorithm

Identifying the close contacts in a big city is a problem of big data. Based on our previous research on big data [11], we propose a fast method to identify the close contacts of epidemiology [4][5]. This method uses traffic data to screen the flow of people. Through clustering hierarchically, the groups of people that have been in close contact with a confirmed case can be identified. This method is from the perspective of the government, and its goal is to improve the monitoring ability of the social health system.

In the method, the government collects people's location data in a city in a day from the location information or the snapshot data of people's trajectories from 8:00 am to 10:00 pm every T hours, with a total of 14/T+1 snapshots (see Figure 1). The concept of snapshot is the premise of this method named multi-snapshot multi-stage minority K-means (3M K-means) algorithm. In each snapshot, a multi-stage minority K-means (2M K-means) is executed. In each stage, minority K-means (1M K-means) is executed.

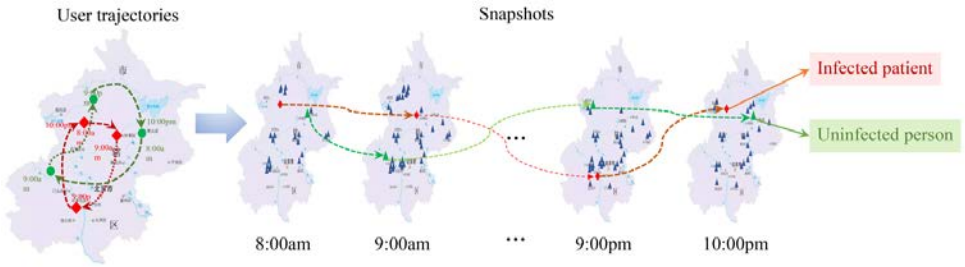

Fig. 1. Schematic of snapshots. The first picture on the left shows the trajectories of users in a day. The four pictures on the right are snapshots of a day showing the location information of users at different times.

The 1M K-means algorithm is a modification of the K-means algorithm. In the process of 1M K-means, if no confirmed case is present in a class during current iteration, then this class is considered to be out of risk and discarded. Thus, the center of this class no longer changes with iteration. If an uninfected point is continuously divided into a class that excludes risk in multiple iterations, then this point can be directly discarded and will not participate in subsequent iterations.

After completing a 1M K-means, most classes and points that exclude the risk of infection are quickly discarded and with much fewer data another 1M K-means in next stage is executed. When the clustering results do not change, the 2M K-means clustering ends, and the points with a close contact risk in this snapshot are obtained. The 3M K-means clustering algorithm can get the infection probability of each close contact by superposing the results of each 2M K-means clustering algorithm. (see Figure 2)

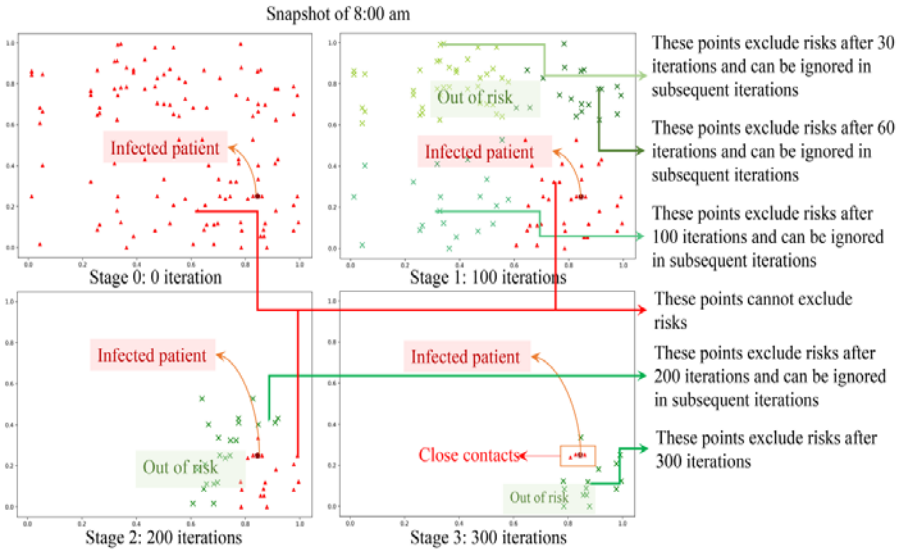

Fig. 2. The iterative process of 2M K-means algorithm for one snapshot at 8:00 am on a particular day. As the number of iterations increases, the points that have been excluded from risk are gradually ignored, and fewer points need to participate in clustering. Points with close contact risk can be obtained after clustering is completed. In the process of 1M K-means, an increasing number of points are ignored in the calculation as the number of iterations increases. In the actual calculation, although the initial points are numerous, a large amount of irrelevant data is rapidly discarded in the early process of iterations, which explains why this algorithm can greatly improve the operation speed.

Within one stage of 2M K-means, the following operation steps of the 1M K-means algorithm are executed:

(1) Before clustering, given K clustering centers, the initial eigenvalue of each cluster is 0, indicating that the risk cannot be excluded.

(2) Clustering is performed by the K-means algorithm. For point m whose eigenvalue is 0, the distance between the point and each class is calculated, and the point is classified into the nearest class. The distance from point m to class n can be expressed as $d = L(m) - L(cn)$, where c represents the center of class n.

(3) If the point with an eigenvalue of 1 does not appear in a certain class in successive p iterations, then this eigenvalue can be rewritten as -1, and the risk of this class is excluded. The class can be discarded directly. In the subsequent iterations, the center position of the class will not change with the iteration.

(4) If a point whose eigenvalue is not 1 continuously appears in the same risk exclusion class in successive q iterations, then the eigenvalue of the point can be marked as -1, that is, the point can exclude risk and can be discarded. This point can be ignored in the subsequent iterations.

(5) Steps 3 and 4 are repeated. When the clustering result remains the same, the 1M K-means of this stage ends.

In each stage of 2M K-means, the 1M K-means algorithm applies the K-means algorithm and constantly reduces the number of participating points, which can greatly improve the operation speed. In the next stage of 2M K-means, the 1M K-means algorithm has much reduced number of data than last stage. This further cuts down the computational time. This 2M K-means algorithm gains additional advantages when the amount of data is large. With the increase in data amount, the 2M K-means algorithm saves more time compared with the ordinary K-means algorithm.

3 Experiments and Discussions

We tested the 2M K-means algorithm and found that for points fewer than 10000, the operation can be completed within one second. When the number of points increases to 5 millions, the operation time does not increase too much, and can be done within 15 minutes. Table 1 shows the running time of the algorithm.

Table 1. Comparison of running time.

Algorithm	Number of points	Number of points
K-means	10	0:00:00.003571
3M K-means	10	0:00:00.000967
K-means	100	0:00:00.040075
3M K-means	100	0:00:00.010462
K-means	1000	0:00:00.430248
3M K-means	1000	0:00:00.093797
K-means	10000	0:00:03.339731
3M K-means	10000	0:00:01.068324
K-means	100000	0:00:46.214205
3M K-means	100000	0:00:11.750244
K-means	1000000	0:12:17.303995
3M K-means	1000000	0:04:48.122983
K-means	5000000	1:19:11.132792
3M K-means	5000000	0:15:51.243982

The 3M algorithm is based on the clustering needs of large-scale data and improves the general clustering algorithm. With a given confirmed case, most points and clusters can be excluded quickly during the clustering process, thereby tremendously improving the calculation efficiency and making the user's location data sufficient for use. Given the limited infectious range of a confirmed case, many of the points can be ignored during this process. Thus, the amount of data involved in the calculation can be reduced to the greatest

extent possible. This algorithm can remarkably improve the operation speed of the information mining algorithm, which can reach about 10 times the speed of the ordinary clustering algorithm for a city with 5 million people. With the increase of data, the speed of this algorithm becomes much faster than that of the ordinary K-means algorithm.

This algorithm reduces the amount of data and considerably improves the speed of clustering by quickly ignoring the excluded risk classes and points in the process in the early stages, whereas traditional algorithms involve with O(N2) computational complexity which needs several days, impracticably for the COVID-19 urgent situations. The 3M algorithm greatly cuts down the computational time, thereof making the rapid warning of close contacts practicable.

4 Conclusions

The 3M algorithm is a “finding a needle in a haystack” cost-saving algorithm because only less than 1/1000 people are diagnosed as infected in a city. Directly calculating people’s trajectories within an acceptable time for a big city is infeasible on account of the big data. This simple, yet efficient, algorithm conspicuously expedites the clustering by nearly 80 times for a city with a population of more than 5 million people, thereby providing the list of close contacts in an acceptable time in practice.

For government agencies, this method are effective anti-epidemic operation tools that can help people resume work with confidence, and hence improving the social health system and public work order in a number of countries. The methods are simple, yet very practicable for the COVID-19 situations. For users, authorized data sharing enables them to have clear and accurate assessments of their action path, take precautions, and make preparations.

This work was supported by the National Social Science Foundation of China (18ZDA309), the National Natural Science Foundation of China (71531012, 62072463), and the Natural Science Foundation of Beijing (4172032).

References

1. Timeline of China releasing information on COVID-19 and advancing international cooperation on epidemic response. China Daily (2020) http://www.chinadaily.com.cn/a/202004/06/WS5e8b2f5aa31012821728496b_3.html .
2. Coronavirus: the first three months as it happened. Nature, 2020. doi: 10.1038/d41586-020-00154-w.
3. Health QR code helps curb the spread of COVID-19, China Daily, http://www.chinadaily.com.cn/a/202003/27/WS5e7dd30da310128217282956.html (2020-03-27).
4. X. Liang, Y. Xue, X. Zhao, A location-based method and system for detecting close contacts of patients with epidemic diagnosis, China Patent Number 202010267024.9 (2020-04-03).
5. X. Liang, Z. Feng, Y. Xue, A multi-level close contact person detection method and system based on successive decreasing clustering, China Patent Number 202011397066.0(2020-12-04)
6. Z. Yue, L. Ma, R. Zhang, Comparison and validation of deep learning models for the diagnosis of pneumonia, Computational Intelligence and Neuroscience, (2020), Article ID 8876798, https://doi.org/10.1155/2020/8876798

7. J. S. Jia, X. Lu, Y. Yuan, et al. Population flow drives spatio-temporal distribution of COVID-19 in China. Nature, (2020). doi: 10.1038/s41586-020-2284-y.
8. C. Sabanayagam, D. Xu, D. S. Ting, et al. A deep learning algorithm to detect chronic kidney disease from retinal photographs in community-based populations. Lancet Digital Health, (2020). doi:10.1016/S2589-7500(20)30063-7.
9. M. Monteiro, V. Newcombe, F. Mathieu, Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head CT using deep learning: an algorithm development and multicentre validation study, Lancet Digital Health, (2020). doi: 10.1016/S2589-7500(20)30085-6.
10. A. Cho, AI systems aim to sniff out coronavirus outbreaks. Science. 368, 6493, 810-811,(2020). doi: 10.1126/science.368.6493.810
11. C. Ma, X. Liang, Y. Ma, A succinct distributive big data clustering algorithm based on local-remote coordination, Proc. IEEE International Conference on Systems, Man and Cybernetics, (2015), 1839-1844, Hong Kong. doi: 10.1109/SMC.2015.322