2020년 한국항해항만학회 추계학술대회 및 ICMASS 2020
2020년 11월 11일(수)~13(금) 온라인 웨비나(한국해양대학교)

A Group Identification Algorithm for Distinguishing Close Contacts in Ships

*Qian-Feng Lin** · † *Joo-Young Son*

**Student, Graduate School of Korea Maritime and Ocean University, Busan 606-791, Korea*
† *(corresponding author)Professor, Division of Marine IT Engineering, Korea Maritime and Ocean University, Busan 606-791, Korea*

Abstract : *There was an outbreak of COVID-19 on the Diamond Princess cruise ship. Distinguishing close contacts is the important problem to be addressed. Close contacts mean people who stays with the patients of disease like COVID-19 over a period of time. The passenger position on board can be obtained by indoor positioning technology. The feature of close contacts is similar location with COVID-19 patients. Therefore, this paper proposed the idea of distinguishing close contacts on board based on DBSCAN algorithm.*

Key words : *COVID-19, Close Contacts, DBSCAN, Cruise Ships*

1. Introduction

The COVID 19 outbreaks on the Diamond Princess cruise ship (2666 passengers, 1045 crew; total 3711) resulted in 712 infected persons, or about 20% of the ship's population. Numerous sources have suggested that quarantine measures on the Diamond Princess. This ship docked off Yokohama, Japan, on February 4, 2020. But they could not control a COVID 19 outbreak. Because the condition of ships allowed spread of the virus among passengers[1]. Therefore, distinguishing close contacts in ships is the most important problem in reducing the risk of transmission and protecting the health of passengers.

2. Previous works

Different starting points and criteria usually lead to different taxonomies of clustering algorithms. The k-means is one of the simplest and popular unsupervised machine learning algorithms. The k-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster. The k-means performs iterative calculations to optimize the positions of the centroids[2].

Hierarchical clustering is an algorithm that groups similar objects into groups called clusters. Hierarchical clustering starts by treating each observation as a separate cluster. It identifies the two clusters that are closest together and merges the two most similar clusters. This process continues until all the clusters are merged together[2].

The grid-based clustering is particularly appropriate to deal with massive datasets. The principle is to first summarize the dataset with a grid representation, and then to merge grid cells in order to obtain clusters[3].

The mean shift clustering is a centroid-based algorithm. It works by updating candidates for centroids to be the mean of the points within a given region. These candidates are filtered to eliminate near-duplicates to form the final set of centroids[3].

Density-based spatial clustering of applications with noise (DBSCAN) is a well-known data clustering algorithm that is commonly used in data mining and machine learning. The key idea of the DBSCAN is that a cluster in data space is a contiguous region of high point density. It is separated from other such clusters by contiguous regions of low point density[4].

The k-means is easy to implement. It may be computationally faster than hierarchical clustering when the number of clusters is small. But the number of clusters is difficult to predict. Hierarchical clustering outputs a hierarchy. It provides more informative than the unstructured set of clusters returned by the k-means. The disadvantage of hierarchical clustering is that the order of the data has an impact on the final results. The grid-based clustering is its significant reduction of the computational complexity. Meanwhile, the grid-based clustering needs a large memory to store the data. The mean-shift makes no model assumptions, unlike the k-means. The mean-shift only sets one parameter, which determines automatically the number of clusters. The mean-shift algorithm, however, does not work well in case of high dimension, where the number of clusters changes abruptly[4].

The DBSCAN does not require to specify the number of

clusters. The DBSCAN is also easy to implement. In addition, it is possible to work well in high dimension data. The order of the data does not have an impact on the DBSCAN. More importantly, the DBSCAN can find arbitrarily-shaped clusters surrounded by a different cluster. Actually, in our idea, the cluster is made up of the location of close contacts. The shape of clusters is arbitrarily shaped. Therefore, the DBSCAN is the most appropriate to apply to our model.

3. Distinguishing close contacts in ships

The ship indoor environment is the narrow space. The precise positions of passengers are needed because of the complex ship environments. However, the reflection and scattering of signals can easily occur in the internal environment of a ship, which may increase the difficulty of getting passenger location in ships. In addition, the passenger activities are restricted and their behavior patterns are quite different from those in the ordinary building environment. The data mining and machine learning technologies are essential to identify groups. The ship environment is complex and changeful. Therefore, a machine learning should be applied to rules from complex ship environments. Epsilon and MinPoints are important to the DBSCAN algorithm. Because two parameters are not stable due to the changeful passenger activities and ship environment. Hence, two parameters need to be adjusted by a machine learning technology to adapt to the different and complex ship environment (Figure 1).

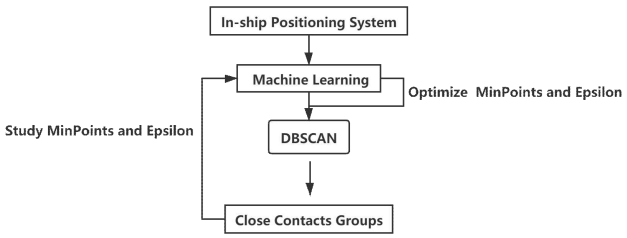

Figure. 1 The flow chart of our proposed method

The patients of infected by disease like COVID-19 and close contacts to them can be observed as one group. The DBSCAN groups together points that are close to each other based on a distance measurement (usually Euclidean Distance) and a minimum number of points. It also marks as outliers the points that are in low-density regions.

The DBSCAN is proposed to identify the close contacts. The input data are set to the passenger positions for the past 14 days. Firstly, the patient and other passenger positions are fed into the DBSCAN algorithm. Secondly, the two parameters are generated by defining the formula. Two parameters are derived by a machine learning algorithm. Finally, the close contacts can be distinguished according to the passenger name. (Figure 2).

4. Conclusion

Isolation of close contacts on board ships is an effective way to reduce the risk of the virus spreading. The detection of close contacts has become one of the most important issues for the protection of the disease infection. The DBSCAN can identify the close contacts based on their position similarity to patients.

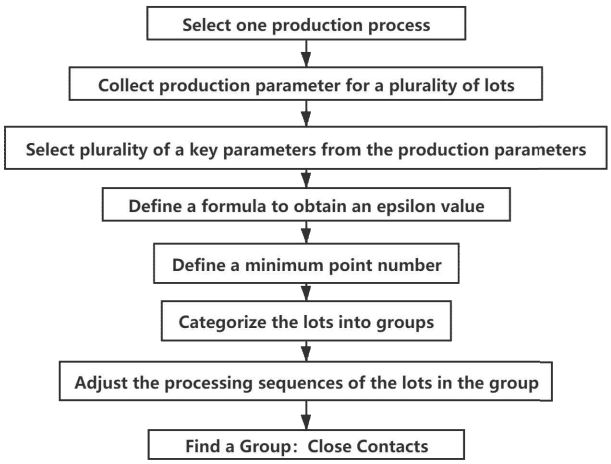

Figure. 2 The flow chart of distinguishing close contacts based on the DBSCAN

The patients and close contacts to them should be observed as one group. But people behave differently depending on their environment. The ship interior has narrow spaces, and the human activities themselves are restricted by the ship environments. The shape of the group is very easy to be changed according to the different environments. Therefore, the minPoints and Epsilon of the DBSCAN should be generated dynamically in order to ensure the accuracy of clustering the groups to distinguish the close contacts. In the future, these two parameters will be obtained in combination with machine learning technologies.

References

[1] COVID 19 outbreak on the Diamond Princess Cruise Ship in

February 2020, https://onlinelibrary.wiley.com/doi/full/10.1002/jgf2.326
[2] A. k. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," ACM computing surveys (CSUR), Vol. 3, pp. 264-323, 1999.
[3] P. Berkhin, "A survey of clustering data mining techniques," Grouping multidimensional data. Springer, Vol. 1, pp. 25-71, 2006.
[4] R. Xu, and W. Donald, "Survey of clustering algorithms," IEEE Transactions on neural networks, Vol. 3, pp. 645-678, 2005.