Comparative analysis of clustering algorithm and improved algorithm application in shipping congestion

Wang Liyang

Jilin University, Software Institute, Changchun, Jilin province, 130012, China

Wangly0618@outlook.com

Abstract. In recent years, due to the lack of reasonable planning of ship operation routes, shipping congestion accidents have been increasing, seriously restricting port development and channel operation, posing a great threat to ship navigation safety and greatly restricting the development prospects of the shipping industry. This paper aims to compare and analyze the application of the clustering algorithm and its improved algorithm in solving port and waterway congestion problems. By analyzing the literature on the use of clustering algorithms to solve shipping congestion problems, and comparing the advantages and disadvantages of multiple clustering algorithms in solving problems, this paper concludes that Partition-based Methods are more suitable for identifying port congestion and the improved Fuzzy DBSCAN algorithm for identifying channel congestion. The research in this paper will help to select clustering algorithms when solving shipping congestion problems in the future.

Keywords: Channel Congestion, Port Congestion, Clustering Algorithm, DBSCAN, K-Means.

1. Introduction

Shipping congestion is a significant concern when considering the economy and efficiency. It leads to extended vessel waiting times, and reduced service quality, and adversely affects competitiveness and demand [1]. Approximately 90% of global trade volume is estimated to be conducted through maritime routes, emphasizing the pivotal role of ports and waterways in supporting economic growth [2]. Most scholars focus on analyzing the causes of shipping congestion and putting forward countermeasures and improvement suggestions according to the analysis results, among various data mining methods, cluster analysis contributes significantly to the research of maritime traffic safety. Shipping congestion mainly includes port congestion and channel congestion. For the problem of port congestion, some scholars have proposed different port congestion indicators to judge the degree of port congestion, for example, AbuAlhaol et al. introduced three indicators with a focus on "Big Data-Driven" approaches for evaluating marine traffic congestion. These indicators include the density of seaports in a given area, the level of spatial complexity, and the average waiting time for ships [2]. The identification problem of channel congestion area is essentially to find the high-density area of ships in the channel, and then combine the traffic flow index of the high-density area to determine whether the channel is congested, in this problem, density-based cluster analysis, unlike traditional clustering methods, does not require the pre-setting of the number of clusters and has the capability to discover clusters of varying numbers and shapes within datasets that may contain noise. Therefore, it finds widespread application in the investigation of channel congestion issues. The topic of this paper is to study the application of the clustering algorithm in solving the problem of shipping congestion. In order to solve the problem of shipping congestion, this paper starts by comparing and analyzing the effectiveness of the solutions given in the latest literature, draws conclusions, and puts forward predictions for solving the problem of shipping congestion in the future.

2. Clustering Correlation Theory and Algorithm

2.1. Data Clustering Methods

Data clustering methods can be primarily categorized into Partition-based methods, Density-based methods, Hierarchical methods, and so on. In this paper, we will select the most representative algorithm analysis and comparison among the three classifications.

2.2. Classical Clustering Algorithm

2.2.1. K-means clustering algorithm (k-means). The k-means algorithm belongs to Partition-based methods. The basic idea of k-means is to find a division scheme of k clusters through iteration so that the mean of these k clusters is used to represent the corresponding types of samples to minimize the overall error [3]. Its advantages are simple principle, easy implementation, fast convergence speed, and the parameters are only the number of clusters K. The drawback lies in the challenging task of selecting the appropriate value for K, as well as the difficulty in achieving convergence for non-convex datasets. Moreover, the results are limited to local optima and are susceptible to noise and outliers.

2.2.2. Density-Based Spatial Clustering of Applications with Noise algorithm (DBSCAN). The DBSCAN algorithm is one of the most representative Density-based methods [4]. DBSCAN defines clusters as the largest collection of densely connected points. It is capable of dividing areas with sufficient density into clusters, discovering clusters of any shape and count within noisy datasets, and automatically determining the number of clusters while identifying noise points within the dataset. DBSCAN algorithm needs to manually set two parameters in the clustering process: ε (neighborhood radius) and MinPts (neighborhood density threshold), different combinations of parameters have a significant impact on the final clustering results, making it generally challenging to determine these two parameters during the clustering process.

The main advantages of DBSCAN are:

(1) There is no need to manually set the number of clusters expected to be generated by clustering, and clusters of different shapes and sizes in the dataset can be identified, and the clustering results are not disturbed by noise points and do not depend on the order of data input.

(2) Only two parameters need to be set in advance in the algorithm.

But its disadvantages are also obvious:

(1) Not suitable for datasets with concentrated data and uneven density distribution.

(2) It is difficult to manually select suitable parameters, but the quality of clustering results depends on the selection of input parameters.

(3) When the amount of data in the dataset is large, the algorithm runs slower and the clustering convergence time is longer.

2.2.3. Clustering Using Representatives (CURE). The CURE algorithm belongs to the category of Hierarchical Methods, as it combines and balances both point-based and centroid-based approaches [5]. Instead of representing a class by a single centroid, it selects a fixed number of representative points in the data space. Having more than one representative point per class allows CURE to accommodate non-spherical geometries. Class shrinkage or condensation can help control the effects of outliers. As a result, CURE treats outliers more robustly and is able to identify non-spherical and large class sizes [6].

Advantages: (1) Clusters of complex spaces can be discovered; (2) Less affected by noise.

Disadvantages: There are many parameters that cannot be applied to large data sets, and sampling is error-informed.

3. Port Congestion Identification

The first step in determining port congestion is to process AIS data: Calculate the normalized Port Congestion Indicators (PCIs) by using geospatial algorithms: spatial concentration, spatial density, and average service time [7].

Ibrahim AbuAlhaol and Vinícius Barreto Martins et al. used different clustering algorithms to extract useful information. Inspired by Martins's thesis [8], Vinícius Barreto Martins et al used k-means and DBSCAN algorithms to cluster normalized port congestion metrics. The k-means aggregates the periods into three clusters. The DBSCAN sets the MinPts, and after setting the MinPts values, calculates the Best Ept Value (maximum curvature point or maximum effective slope on the graph) [7]. Similar but different, Ibrahim AbuAlhaol used a k-means++ clustering algorithm to characterize the congestion level of ports [2]. The three clustering distances (1,1,1) obtained by the clustering algorithm are, from near to far: highest congestion cluster, moderate congestion cluster, and lowest congestion cluster.

The confusion matrix, precision, recall, and f1-score are the most commonly used metrics [9]. Martins refers to the standard of accuracy as the ratio of the total number of correct predictions and the total number of predictions [10]. The KNN algorithm and the R.Forest algorithm were used for Confusion Matrix based on K-means and Confusion Matrix based on DBSCAN, respectively, the accuracy of the K-means was higher than that based on the DBSCSN. It is concluded that the clustering effect of the K-means is better than that of the DBSCAN in port congestion identification. Meanwhile, Ibrahim AbuAlhaol also got the right results using the K-means++ algorithm.

In the current judgment of port congestion, because no matter what clustering algorithm is used, in the PCIs judgment standard, only 3 clustering results are finally obtained, and the parameters of the segmented clustering algorithm naturally only need the final number of clusters, and the final result of density-based clustering algorithm does not necessarily get 3 clusters, which is not suitable for PCIs, especially DBSCAN algorithm, which has the disadvantage of poor clustering effect on uneven data clustering. While the third PCI, average service time, expressed by the average time required for a ship to enter, load/unload and leave the port AOI, there are large differences between individual ships, resulting in uneven data density, thereby reducing the accuracy of the DBSAN algorithm. Therefore, it is suitable to use the Partition-based Methods to determine port congestion.

4. Channel Congestion Identification

4.1. *Improved Fuzzy DBSCAN algorithm*

The essence of identifying the port congestion problem is to identify high-density concentration areas of ships within the water area, and this step requires the use of clustering algorithms. Subsequently, combined with the traffic flow information, determine whether congestion has occurred. Since the Partition-based Methods parameter requires the number of clusters to be entered, but the required number of clustering clusters cannot be determined because the number of high-density areas is uncertain, the Partition-based Methods are not applicable. Due to the irregular geometry of the channel and the characteristics of the uneven density of the processed AIS dataset, the shortcomings of the classical DBSCAN algorithm led to its low accuracy in identifying high-density accumulation areas of ships, so the classical DBSCAN needs to be improved.

In view of the problems that the parameters of the classical DBSCAN algorithm are difficult to determine and cannot be applied to the density of uneven datasets, Liu Yuan combines the classical DBSCAN algorithm with fuzzy set theory, simulates the approximation of input parameters by defining soft constraints, and determines the degree of each data belonging to each cluster through the membership function, and three fuzzy extensions of the DBSCAN are obtained [11].

The first fuzzy extension is Fuzzy Core DBSCAN, which performs soft constraints on the neighborhood density parameter $MinPts$, using the maximum neighborhood density $Mpts_{\mathrm{Max}}$ and the neighborhood density minimum $Mpts_{\mathrm{Min}}$ of a certain point ε neighborhood density to replace the original $MinPts$ to allow the generation of clusters with fuzzy core points, that is, clusters with different neighborhood density core points. Fuzzy Border DBSCAN, replaces concrete ε by piecewise linear functions defined by $\varepsilon_{\mathrm{Min}}$ and varepsilon $\varepsilon_{\mathrm{Max}}$ as soft constraints to relax constraints on ε to allow the generation of clusters with overlapping boundaries.

Fuzzy DBSCAN includes the previous two fuzzy extensions, with both soft constraints described above, allowing the generation of clusters with fuzzy core points and fuzzy overlapping boundaries. The purpose of using the combination of the above two improvements is to try to reduce the sensitivity of the algorithm to the input parameters $MinPts$ and ε, make the algorithm suitable for datasets with uneven density, and improve the clustering effect of the algorithm. Since the Fuzzy DBSCAN algorithm needs to set four input parameters, the dependence of the algorithm on the input parameters is increased, and the performance does not increase but decreases, so the improved Fuzzy DBSCAN algorithm is improved again on the basis of Fuzzy DBSCAN.

The core idea of the improved Fuzzy DBSCAN algorithm is similar to the Fuzzy DBSCAN clustering process, with the added requirement of calculating the distance between every pair of points, use this value to continuously update $\varepsilon_{\mathrm{Max}}$ so that $\varepsilon_{\mathrm{Max}}$ is equal to the maximum value of the calculated distance between all two points, so that only $\varepsilon_{\mathrm{Min}}$ needs to be specified at the beginning of the clustering, and the algorithm will dynamically generate $\varepsilon_{\mathrm{Max}}$. The Fuzzy DBSCAN algorithm calculates the number of points contained in each point ε neighborhood, and uses this value to continuously update $Mpts_{\mathrm{Max}}$ so that $Mpts_{\mathrm{Max}}$ is equal to the maximum number of points contained in each point ε neighborhood. In this way, only $Mpts_{\mathrm{Min}}$ needs to be specified during the clustering process, and the algorithm will dynamically generate $Mpts_{\mathrm{Max}}$. At the same time, the clustering effect of the algorithm is improved by optimizing the algorithm process and the internal membership degree discrimination criterion, and the improved Fuzzy DBSCAN algorithm only needs to set $\varepsilon_{\mathrm{Min}}$ and $Mpts_{\mathrm{Min}}$ to complete clustering.

The high-density gathering area of ships was identified by selecting the channel in the waters of Wusongkou, and seven high-density areas were clustered by using the improved Fuzzy DBSCAN algorithm, and the areas obtained by combining the high-density areas of traffic flow information were judged to be channel congestion areas. The results indicate the successful identification of traffic congestion areas in the waterway using the improved Fuzzy DBSCAN algorithm.

F-Score is a comprehensive index for evaluating clustering results, in the F-Score standard judgment clustering effect, the clustering effect of improved Fuzzy DBSCAN has been significantly improved, the clustering results have high accuracy, and the improved algorithm reducing the difficulty of selecting parameters, with the advantages of density-based clustering algorithm, can find any shape and number of clusters in the dataset, and has good applicability to datasets with uneven density. Moreover, the clustering effect is better than the classical DBSCAN algorithm and three other fuzzy extensions, with higher operating efficiency and shorter running time, which makes the improved Fuzzy DBSCAN suitable for identifying port congestion. However, as the data volume increases, the run time of the improved Fuzzy DBSCAN algorithm will exponentially increase due to its significantly higher complexity compared to the other four algorithms.

4.2. CURE algorithm

To address the challenge of handling large data sets, the CURE algorithm is unsuitable. However, a stream processing technique has emerged as a solution for monitoring ship trajectories [12]. This technique divides the incoming data into frames and processes each frame comprehensively before the arrival of the next frame. By sequentially feeding the AIS dataset into the CURE algorithm until the final batch of data, this approach allows the CURE algorithm to effectively handle large data sets and adapt to real-time data. This method of data entry removes memory limitations by sacrificing disk space, which inevitably increases the running time of the algorithm. This CURE algorithm can realize the monitoring of the trajectory of ships, and its principle is suitable for identifying channel congestion, but it still has the problem of too many parameters and does not completely solve the problem of difficulty in processing large data sets, which is the direction of future improvement. Due to its shortcomings, it is inferior to the improved Fuzzy DBSCAN algorithm in identifying channel congestion.

5. Conclusion

In this paper, a variety of clustering algorithms currently applied to shipping congestion are analyzed and compared. Because Partition-based Methods, such as the k-means, whose parameter is the number of clusters, conform to the classification rules of PCIs and obtain high accuracy in the test, it is suitable for solving port congestion problems. The improved Fuzzy DBSCSN algorithm simplifies the parameters, solves the problem of complex parameters of the classical DBSCAN algorithm, and compares the multiple shortcomings of the CURE algorithm, showing that it is more suitable for solving the problem of channel congestion. Based on the above two aspects, the application of the clustering algorithm and improved algorithm in shipping congestion is described.

If the improved algorithm of CURE algorithm can solve the current problem, its advantages can ensure that it will play a greater role in solving channel congestion. Since only some clustering algorithms and improved algorithms are currently used to solve the shipping congestion problem, this paper is only a phased analysis, and more clustering algorithms may be applied to deal with shipping congestion in the future.

References

[1] SAEED, N. et al. Governance mode for port congestion mitigation: A transaction cost perspective. NETNOMICS: Economic Research and Electronic Networking, v. 19, n. 3, pp. 159-178, 2018.

[2] AbuAlhaol, R. Falcon, R. Abielmona and E. Petriu, "Mining Port Congestion Indicators from Big AIS Data," 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 2018, pp. 1-8, doi: 10.1109/IJCNN.2018.8489187.

[3] X. Chu, J. Lei, X. Liu and Z. Wang, "KMEANS Algorithm Clustering for Massive AIS Data Based on the Spark Platform," 2020 5th International Conference on Control, Robotics and Cybernetics (CRC), Wuhan, China, 2020, pp. 36-39, doi: 10.1109/CRC51253.2020.9253451.

[4] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//kdd. 1996, 96(34): 226-231.

[5] P. Bharadwaj, R. Gupta, R. Gurjar and A. Singh, "Importance of CURE Clustering Algorithm over K-Means Clustering Algorithm for Large Data-set," 2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 2023, pp. 421-426, doi: 10.1109/ICSCCC58608.2023.10177015.

[6] Guha S , Rastogi R , Shim K .CURE : An Efficient Clustering Algorithm for Large Databases[J].Information Systems, 1998, 26(1):35-58.DOI:10.1016/S0306-4379(01)00008-4.

[7] Martins, Vinícius & Fernandes Ramos, Ramiro & F. S. Cepeda, Maricruz & Caprace, Jean. (2021). Ocean Port Congestion Indicators - A Machine Learning Approach. 10.17648/sobena-hidroviario-2021-137502.

[8] MARTINS, V. Dynamic Port Congestion Indicators - Case Study of the Ports of Rio de Janeiro and Santos. Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Oceânica, COPPE, da Universidade Federal do Rio de Janeiro. Maio de 2021.

[9] ZOLTAN, C. KNN in Python. https://towardsdatascience.com/knn-in-python. Accessed: 25/01/2021.

[10] FOR GEEKS, G. Confusion Matrix in Machine Learning. https://www.geeksforgeeks.org/confusion-matrix Accessed:05/03/2021.

[11] Liu Yuan. Identification of channel congestion status based on improved Fuzzy DBSCAN[D].Dalian Maritime University,2022.DOI:10.26989/d.cnki.gdlhu.2022.001216.

[12] C. Manyfield-Donald, T. A. Kwembe and J. -R. C. Cheng, "A Modified Clustering Using Representatives to Enhance and Optimize Tracking and Monitoring of Maritime Traffic in Real-time Using Automatic Identification System Data," 2021 International Conference on

Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2021, pp. 285-289, doi: 10.1109/CSCI54926.2021.00119.