Identification of Noise Points in Buoy Telemetry Position Data Based on PSO-DBSCAN Algorithm

Shao Jinxing[1], Zhang Xinliang[2], Xu Liangkun[2*]1 Xiamen Aid to Navigation Department of Donghai Navigation Safety Administration, 361000 Xiamen, China

2 Navigation College of Jimei University, 361021 Xiamen, China

Abstract: Buoys are artificial signs that guide the navigation of ships, and are of great significance to ensuring the safety of ships. The effective use of the buoy telemetry position data can analyze the offset characteristics of the buoy and improve the buoy's navigation aid efficiency. However, due to the influence of data transmission or human factors, the buoy telemetry data often contains noise points, which affects the application of buoy telemetry position data. To identify noise points in buoy telemetry position data, a DBSCAN algorithm optimized by particle swarm optimization is proposed, and the input parameters are determined adaptively to realize the identification of noise points in the buoy telemetry position data. Experiments on the relative buoy telemetry position data in Xiamen Port show that the PSO-DBSCAN algorithm can accurately identify the noise points in the buoy telemetry position data, and its effect is consistent with the actual situation.

Keywords: Buoy; Position Data; Particle Swarm Algorithm; Noise Point Identification

1 Introduction

Buoys are water surface-navigation aids that indicate the range of the fairway, indicate shoals and obstructions, or express special purposes. They are also artificial signs to guide the navigation of ships, which are of great significance to ensuring the safety of ships. Therefore, it is one of the main responsibilities of the navigation security department to ensure the accurate placement of the buoys and to provide the most accurate position information for the ship's navigation [1-2]. The telemetry data of the buoy contains important parameters of the buoy's position, time, voltage, and other operating states of the buoy. It is an important spatiotemporal data type, by analyzing the position data, the similar characteristics of the displacement trajectories of each light buoy can be obtained, and a meaningful displacement pattern can be found. However, the original buoy telemetry data may have errors in manual input, information transmission, sensor data collection, storage, and other links. Taking the telemetry position data of the Xiamen Port 10# buoy from January to June 2021 as an example, there are abnormal positions deviating from the buoy, as shown in Figure 1.

Abnormal buoy position data is a big interference to the utilization of buoy telemetry data. Therefore, these abnormal or noisy data must be preprocessed before using the buoy telemetry position data, and the abnormal or noisy data points of each buoy must be eliminated, and then a high-quality data set can be provided for subsequent use.

Correspondence: 15980991169@163.com

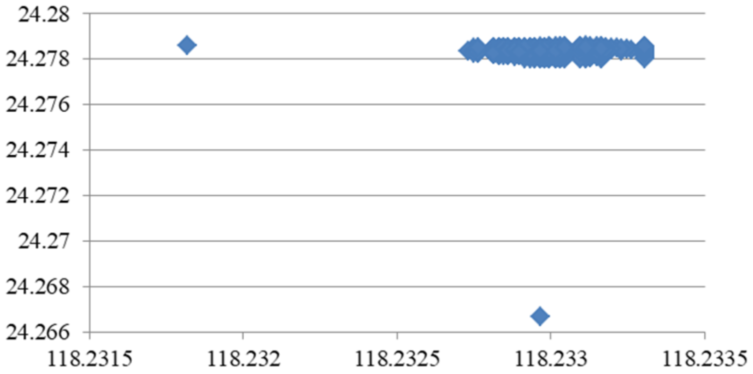

Figure 1 Schematic diagram of an abnormal point of Xiamen Port 10# telemetry position data

2. PSO-DBSCAN Clustering Algorithm

2.1 Basic Idea

The algorithm based on density clustering is a typical unsupervised learning method. It mainly divides the sample points into the data set according to the density characteristics of the data set distribution and has the characteristics of less prior knowledge requirements. Not only that, but density clustering-based algorithms are also able to identify clusters of various shapes and sizes in noisy datasets. As one of the most widely used density clustering algorithms, the DBSCAN algorithm needs to input the neighborhood distance value Eps and the number of sample points in the neighborhood Minpts, and the clustering results are more sensitive to the Eps and Minpts parameters. At present, the improvement of the clustering effect also relies on manual experience to adjust the parameters many times. In addition, the robustness of the DBSCAN algorithm is poor. After changing the target dataset for clustering, the parameters of Eps and Minpts need to be adjusted[3-4]. The particle swarm optimization algorithm is intelligent. It uses the population method to search by simulating the flight and foraging behavior of birds. It can search for the global optimal solution in the solution space of the objective function to be optimized and find the DBSCAN input parameters Eps and the optimal combination of Minpts can realize the adaptive clustering of the DBSCAN algorithm and avoid the influence of the input parameters on the clustering results. Therefore, this paper applies the particle swarm optimization algorithm to optimize the DBSCAN algorithm and proposes a PSO-DBCSAN algorithm based on the particle swarm optimization algorithm to optimize the parameters of the DBSCAN algorithm. The algorithm specifies the number of clusters in the target dataset, and the silhouette coefficient is used as the fitness function of the particle swarm algorithm, the optimal solution is selected in combination with the number of clusters, and finally, the optimal Eps and Minpts parameters are optimized through iteration.

2.2 Parameter range determination

In this paper, the distribution characteristics of the data set are used to determine the range of parameters required by the particle swarm optimization algorithm. First, the distance matrix between sample points is calculated, and then the elements in each column of the distance matrix are sorted in ascending order. After sorting, the second nearest neighbor distance with the size second only to itself is selected and integrated into a set, which is used as the value range of Eps. After determining the value range of Eps, select the maximum value and the minimum value in the Eps set, respectively adopt the mathematical expectation method in formula (1), and finally determine the parameter range of Minpts. In formula (1), n represents the number of samples in the dataset, and Pi represents the number of samples of object i within the range of the maximum Eps parameter.

$$Minpts = \frac{1}{n}\sum_{i=1}^{n} P_i \tag{1}$$

2.3 Fitness function selection

The fitness function, also known as the objective function, is the key to the optimization of the parameters of the PSO-DBSCAN algorithm. The optimization process of the algorithm in this paper mainly uses the clustering metric evaluation function to calculate the fitness of the individual particle swarm. In the absence of sufficient prior knowledge, the main methods used are Davies-Bouldin Index (DBI), Silhouette Coefficient (S(i)), and CH score (Calinski Harabasz Score, CH). Since the DBI index has a poor evaluation effect on the circularly distributed dataset, the CH score will score higher for clusters with convex features than for other types of clusters. Therefore, this paper selects the silhouette coefficient as the fitness function of this paper to optimize the target parameters[5]. Although the adjustment of the position information of the particle swarm itself is the process of seeking the optimal solution for the fitness function, the optimal solution is negatively correlated with the fitness. However, the value range of the contour coefficient is [-1, 1], so the negative number of the contour coefficient can be selected as the fitness function. The mathematical expression of the silhouette coefficient is shown in formula (2).

$$S(i) = -\frac{b(i)-a(i)}{\max\left\{a(i), b(i)\right\}} \tag{2}$$

In the formula, $a(i)$ represents the average distance between the ith object and other objects in the cluster, and $b(i)$ represents the average distance between the i-th object and the objects in other clusters except the cluster where i is located. $S(i) \in \left[-1,1\right]$, when b is closer to -1, it means that the clustering quality is higher.

2.4 Algorithm Design

The algorithm in this paper sets the dimension search space to 2, the maximum number of iterations is T, and the fitness function is set to the clustering evaluation function Silhouette. The algorithm implementation steps are as follows:

Step 1: Process the buoy dataset to generate a two-dimensional coordinate dataset.

Step 2: Enter the number of sample clusters for optimal clustering (best_n).

Step 3: Set the two input parameters of the DBSCAN algorithm (the neighborhood distance value Eps, the number of sample points in the neighborhood Minpts) as the optimization goal, and use the method proposed in Section 3.3.2 to determine the range of Eps and Minpts parameters, set The particle dimension is 2, the number of particles is 200, and the maximum number of iterations is 50. Finally, initialize the particle swarm algorithm.

Step 4: Calculate the negative silhouette coefficient of each particle as a fitness value for optimization according to formula (2), and determine the current global extreme value (best_S) and the parameters under the global extreme value (best_Eps, best_Minpts).

Step 5: Update the velocity and position of each particle.

Step 6: Calculate the fitness value (Fitness_i) and the number of clusters (clustering_n) of the particle, and update the individual extreme value (best_i) and the global extreme value (best_S) respectively.

Step 7: Re-initialize the position randomly for some particles that jump out of the set parameter range during the iteration process.

Step 8: Determine whether the number of iterations is reached or whether the fitness of all particles converges to the global extremum (best_S). If the number of iterations T is reached or the fitness values of all particles converge to best_S, return the parameters of the global extremum (best_Eps and best_Minpts) and the global extreme value (best_S); otherwise, return to step 5.

Step 9: Use the optimal parameters to construct the optimal clustering model.

3 Buoy Noise Point Identification Based on PSO-DBSCAN Algorithm

3.1 Buoy Data

Select the 3#, 4#, 5#, 6#, 22A#, 23#, 24#, 35#, 36#, 37#, and 702# buoys and other telemetry data in different periods in Xiamen Port to verify the PSO-DBSCAN algorithm in the buoy. The effect of telemetry location data in noise point detection. Among them, buoys 3#-6# in the main channel of Xiamen Port use data from April to June 2020, 22A#, 23# and 24 floats use data from January to March 2021, 35#, 36#, 37# And the 702# buoy selects the data from July to September 2021. The telemetry position data for these buoys are shown in Figures 2-5.

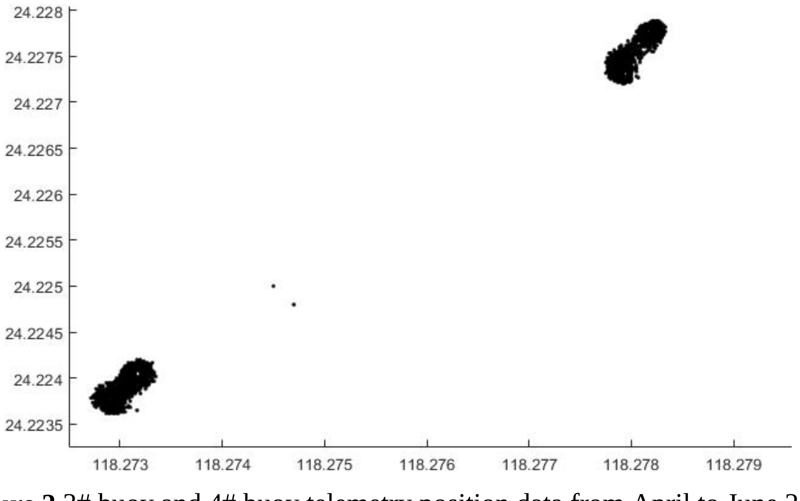

Figure 2 3# buoy and 4# buoy telemetry position data from April to June 2021

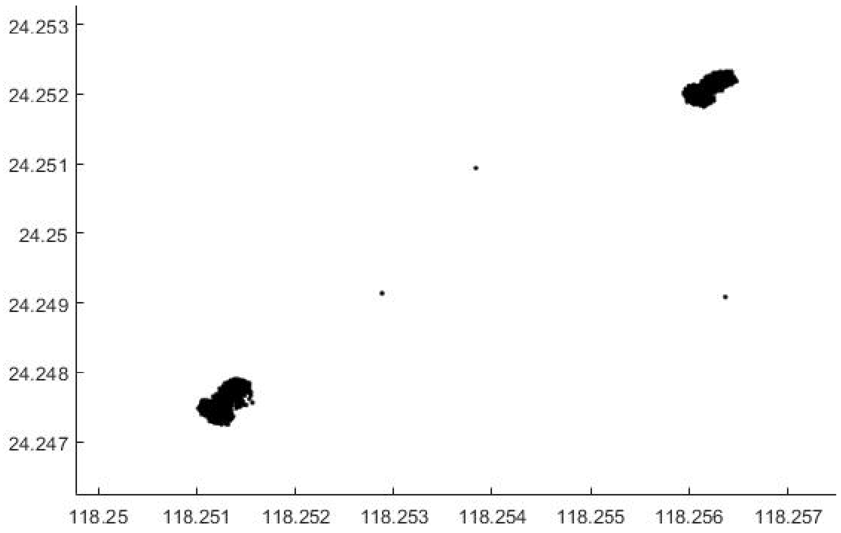

Figure 3 5# buoy and 6# buoy telemetry position data from April to June 2021

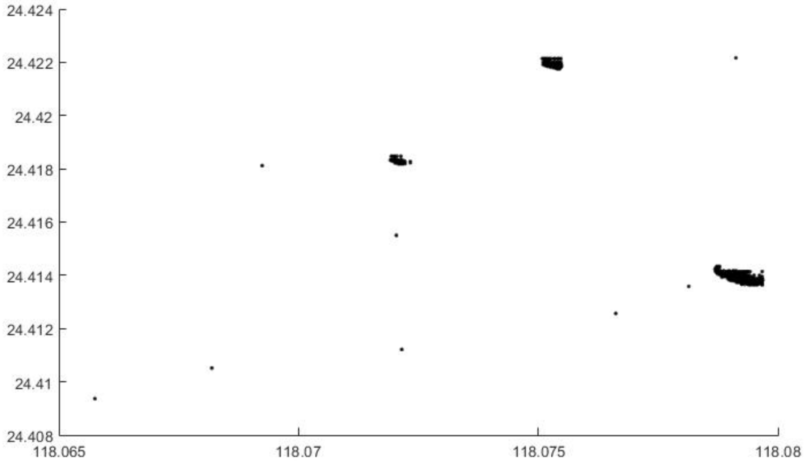

Figure 4 22A# buoy, 23# buoy, and 24# buoy data from January to March

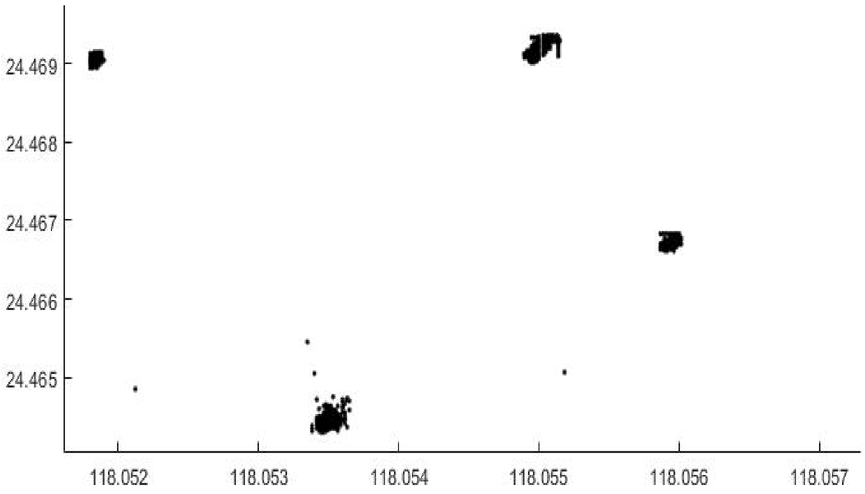

Figure 5 35# buoy, 36# buoy, and 37# buoy data from July to September

3.2 Preliminary processing of buoy data

In some buoy position data, the longitude or latitude value of some buoy telemetry position data deviates from other position points, such data can be deleted directly. Taking the main channel 3# float as an example, the longitude and latitude data at 6:03 on March 7, 2021, obviously deviates from other positions too much (as shown in Table 1). For such data, it can be deleted directly.

Table 1 Partial telemetry position data of the 3# buoy in the main channel of Xiamen Port

Buoy Number	longitude(E)	Latitude(N)	Time
3 #	118.0753667	24.49998333	2021/03/07 06:03
3 #	118.2782833	24.22781667	2021/04/07 07:03
3 #	118.2782833	24.22780000	2021/05/07 08:03
3 #	118.2452500	24.25738333	2021/06/07 09:03
3 #	118.2782500	24.22771667	2021/07/07 10:03
3 #	118.2782833	24.22768333	2021/08/07 11:03

3.3 Buoy position noise point data identification based on PSO-DBSCAN algorithm

Preliminary processing of the buoy raw data is carried out manually, and only those abnormal positions that are far away from the buoy sinking can be selected, while the noise points located near the buoy sinking are very difficult when the amount of data is large. It is difficult to delete manually. Therefore, after the preliminary processing of the buoy telemetry position data, the PSO-DBSCAN algorithm proposed in this paper is used to identify the noise point data, and the identification effect is shown in Figures 6-9. In the process of identifying noise locations, the PSO-DBSCAN algorithm only needs to specify the number of buoys and determines the parameters of the PSO-DBSCAN algorithm through the strong global optimization performance of the particle swarm algorithm, and automatically realizes the identification of noise locations.

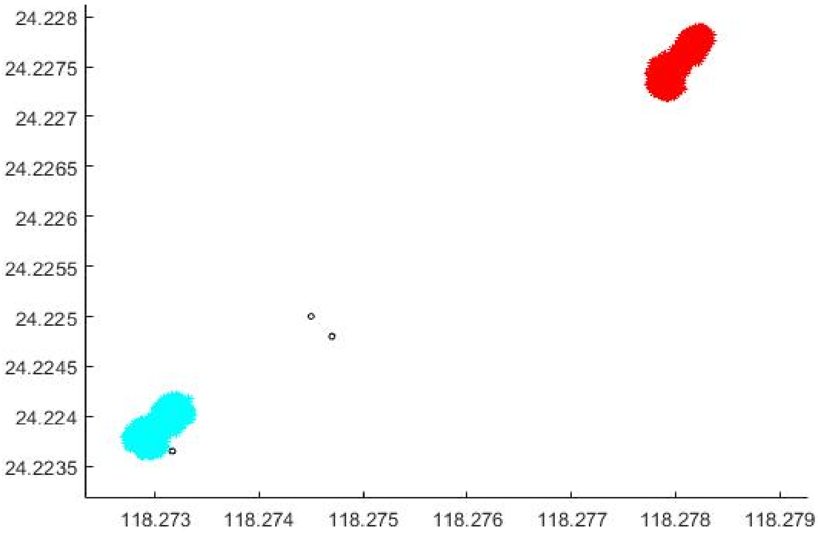

Figure 6 3# buoy and 4# buoy position data noise point recognition effect

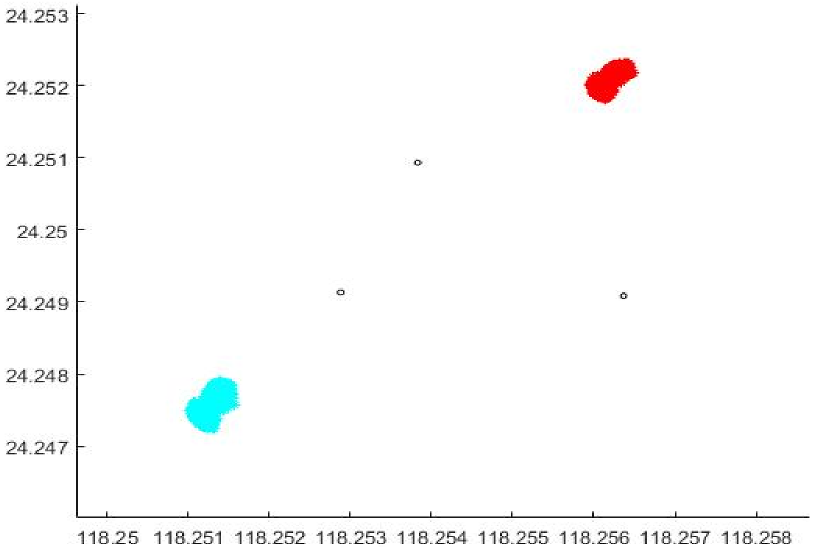

Figure 7 5# buoy and 6# buoy position data noise point recognition effect

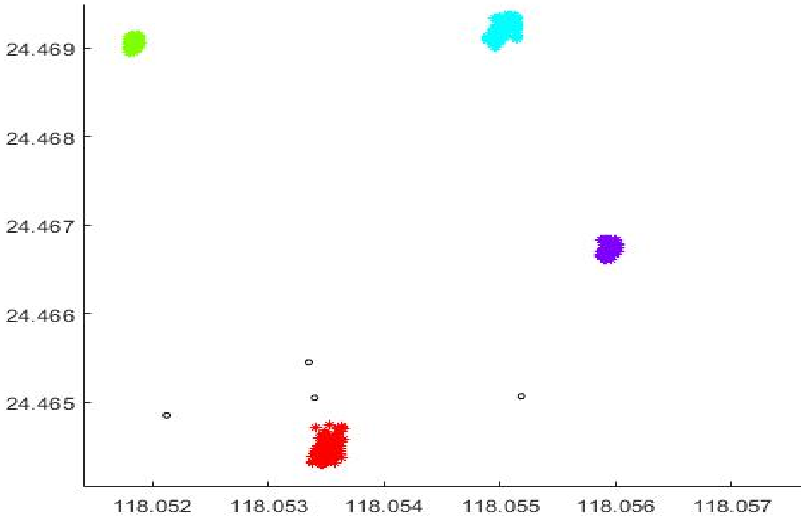

Figure 8 35# buoy, 36# buoy, 37# buoy, and 702# buoy position data noise point recognition effect

The water depth of the main channel of Xiamen Port and its surrounding waters is generally less than 20m, and the total length of the buoy anchor chain is about 80m. Therefore, in general, the point where the distance between the buoy telemetry position data and the sinking rock is greater than 80m is a noise point. It can be seen from Figure 6-Figure 9 that the position points far from these buoys are identified as noise points, which indicates that the noise points of the buoy position data identified by the PSO-DBSCAN algorithm are consistent with the actual situation, and the application effect is good.

4 Conclusion

There may be errors in the buoy telemetry data in manual input, information transmission, sensor data collection, storage, etc. Therefore, before the buoy telemetry position data is used, the data must be preprocessed to provide high-quality data sets for subsequent use. Aiming at the characteristics of buoy telemetry position data, this paper proposes a DBSCAN algorithm PSO-DBSCAN based on particle swarm optimization and applies the algorithm to the identification of noise points in buoy telemetry position data in Xiamen Port, the results show that the algorithm can accurately identify the noise points in the buoy telemetry position data, and its effect is consistent with the actual situation.

References

[1] PAN ZM, SHAO JX, TANG JH, et al. Research on Offset Characteristics of Light Buoys in Main Channel of Xiamen Port[J]. World Shipping, 2020, 43(7): 35-39.

[2] PENG TT. Buoy Position Monitoring and Monitoring Data Processing[J]. Mechanical & Electrical Engineering Technology, 2011, 40(7): 63-65.

[3] Sunita, J.; Parag K. Algorithm to determine ε -distance parameter in density based clustering. EXPERT SYST APPL, 2014, 41: 2939-2946.

[4] Cassisi, C.; Ferro, A.; Giugno, R.; Pigola, G.; Pulvirenti, A. Enhancing density-based clustering: Parameter reduction and outlier detection. Inf. Syst. 2013, 38: 317-330.

[5]Matsumoto, N.; Hamakawa, Y.; Tatsumura, K.; et al. Distance-based clustering using QUBO formulations. Sci Rep.2022, 12: 2669.