ISSN: 2710-0588

Research on Anomaly Detection of Ship Behavior Based on Data Mining

Jun Zhang, Enze Wu, Taizhi Lv

College of Information Engineering, Jiangsu Maritime Institute, Jiangsu Nanjing, 211170, China

Abstract

With the increase in the number of ships and the intensifying maritime traffic, maritime accidents such as ship collisions and grounding occur frequently, posing serious threats to people's lives and property. Therefore, timely detection and correction of abnormal ship behaviors are of great significance for improving navigation safety. The Automatic Identification System (AIS) can provide real-time information on ships' positions, speeds, headings, and more, offering a rich data resource for data mining. Through the analysis of these data, it is possible to deeply explore the navigation patterns and characteristics of abnormal behaviors of ships. This paper applies data mining technology to the detection of abnormal ship behaviors, which not only improves the accuracy and efficiency of detection but also provides maritime regulatory authorities with richer decision-making support.

Keywords

Data Mining, Hive Data Warehouse, AIS Data, Analysis Of Abnormal Ship Behavior.

1. Introduction

With the rapid growth of the global shipping industry, both the number of ships and the complexity of navigation are continuously increasing, making the detection of abnormal ship behaviors particularly critical and urgent. Abnormal ship behaviors not only may trigger maritime accidents, posing threats to personnel safety, but also can cause irreversible damage to the environment and even endanger national marine security [1]. Therefore, accurately and effectively identifying abnormal ship behaviors is essential for enhancing shipping safety and promoting the sustainable development of the marine economy.

As an efficient means of data processing and analysis, data mining technology can extract valuable information and patterns from massive data, providing a new solution for monitoring abnormal ship behaviors [2]. By using data mining methods, we can conduct in-depth research on ship navigation data, observe their dynamic characteristics and behavioral habits, and establish a model that accurately describes the daily behaviors of ships based on these research findings. When a ship's behavior deviates from the normal model, it can be quickly identified and warned, thereby providing effective decision-making support for shipping management agencies.

This paper follows the steps of data collection and preprocessing, feature extraction and selection, model construction and training, and abnormal behavior detection. Data collection and preprocessing involve gathering real-time data from the AIS system, including ship positions, speeds, headings, etc. The data is then cleaned, denoised, and standardized to ensure its quality and accuracy. Feature extraction and selection aim to extract features related to abnormal behaviors from ship data, such as position offsets, speed changes, heading changes, etc. Through correlation analysis, principal component analysis, and other methods, the most representative features are selected as inputs for the model. Model construction and training

ISSN: 2710-0588

mainly utilize data mining techniques, such as cluster analysis and classification algorithms, to build a detection model for abnormal ship behaviors. Historical data is used to train the model and adjust its parameters, improving the model's accuracy and generalization ability. Abnormal behavior detection involves inputting AIS data into the trained model to monitor the ship's navigation status. First, AIS data from ships is collected, and then the motion features and behavior patterns of ships are analyzed and extracted. Next, data mining techniques such as cluster analysis and association rule mining are used to construct a ship behavior model, accurately portraying the normal behaviors of ships. Based on this, an abnormal detection algorithm is further designed to identify abnormal behaviors by comparing actual ship behaviors with the normal behavior model. The data mining-based method for detecting abnormal ship behaviors has high accuracy and reliability, enabling timely detection and warning of abnormal ship behaviors, thereby providing decision-making support for shipping management departments.

2. Research Status of Ship Abnormal Behaviors

2.1. Definition of Ship Abnormal Behaviors

Ship abnormal behaviors generally refer to behaviors that deviate from the usual performance of similar ships in terms of track, speed, or equipment status under specific conditions such as region, weather, hydrology, time, regulations, relationships with nearby ships, and previous performance [3]. Such behaviors may be indicators of potential risks or manifestations of illegal activities. Although ships navigate under human control, ship behaviors are not entirely reflective of human intentions. This is because the marine environment differs from the land, and ships are susceptible to influences such as currents, winds, and waves when navigating at sea. Depending on the usage context, abnormalities have various interpretations. However, there is currently no established standard for defining ship abnormal behaviors. It is known that ship movement abnormalities can manifest in various forms, such as significant deviations from established routes or two ships being sufficiently close to facilitate smuggling. From the perspective of daily concerns of maritime regulatory personnel, ship abnormal behaviors encompass abnormalities in navigation characteristics and navigation activities. Abnormalities in navigation characteristics consider whether motion parameters such as ship speed, position, and trajectory fall within normal ranges, examples include excessive speed, navigation outside designated channels, U-turns within channels, and illegal anchoring. Abnormalities in navigation activities consider whether ships are engaged in illegal activities, such as overloading, illegal operations, and stowaways.

2.2. Classic Methods for Ship Abnormal Detection

Classic abnormal detection methods include clustering-based, statistical, deep learning, classification model-based, and deviation-based approaches. In today's data-driven world, abnormal detection technology plays a crucial role.

(1) clustering-based methods

Normal data instances belong to a cluster within the data, while abnormal data instances do not belong to any cluster [4]. Normal data instances are close to their nearest cluster centroid, whereas abnormal data is far from its nearest cluster centroid. Normal data instances belong to large and dense clusters, while abnormal data instances either belong to small clusters or sparse clusters. By categorizing data into different clusters, abnormal data refers to those that belong to small clusters, do not belong to any cluster, or are far from the cluster center. Examples of methods that identify small clusters as abnormal points include K-means clustering. Methods that identify data not belonging to any cluster as abnormal include DBSCAN, ROCK, and SNN clustering. This approach is suitable for scenarios where labeled data is absent

ISSN: 2710-0588

and can be used to discover unknown patterns or structures in the data. For instance, clustering ship trajectories can categorize normal and abnormal navigation trajectories.

(2) statistical methods

Statistical methods rely on the assumption that the dataset follows a certain distribution or probability model [5]. By determining whether a data point fits this distribution model, abnormal detection is achieved. These methods are suitable for scenarios where data characteristics are relatively stable and conform to specific distributions. Based on probability models, they can be divided into parametric and non-parametric methods. Parametric methods estimate model parameters from data with known distributions. When the difference between a data point and the mean exceeds a threshold, it is considered abnormal. Non-parametric methods, used when the data distribution is unknown, involve plotting histograms to detect whether data falls within the histogram generated by the training set for abnormal detection. For example, by statistically analyzing ship navigation and speed in a certain area, abnormal navigation behaviors of ships can be identified.

(3) neural network-based methods

Neural network-based abnormal detection methods leverage the powerful feature extraction and pattern recognition capabilities of neural networks to identify abnormal patterns in data. Representative methods include Autoencoder (AE) and Long Short-Term Memory (LSTM) neural networks. LSTM can be used for abnormal detection in time series data, detecting abnormal points with significant deviations from predicted values [6].

3. Ship Anomaly Behavior Analysis

3.1. Data Collection and Preprocessing

(1) data collection

The experimental samples for this paper are drawn from a dataset in the Hive data warehouse, specifically from five days of data spanning from June 22, 2023, to June 26, 2023. This time period covers 6,467,283 entries of shipping data, encompassing various types of civilian vessels. These data truly and effectively reflect the dynamic information of ships, including core basic information such as the ships' latitude and longitude, message reception time, and ground speed. The core code for data extraction is as follows.

```
conn = hive.Connection(host=host_name, port=port, username=user)
cursor = conn.cursor()
cursor.execute(“SELECT * FROM ais where receive>=’2023-06-22 00:00:00
        and receive<=‘2023-06-26 23:59:59‘)
df = read_sql(cursor.fetchall(), conn)
```

The output is shown in Figure 1, and it includes fields such as the ship's MMSI (Maritime Mobile Service Identity) code, latitude and longitude, speed, heading, course over ground (COG), reception time, etc.

ISSN: 2710-0588

```
                          receive       mmsi         lng  ...  head  rate  second
0        2023-06-22 01:04:22.125000  413422880  120.842450  ...   259     0     NaN
1        2023-06-22 01:04:22.197000  413819288  118.963780  ...   511  -128    60.0
2        2023-06-22 01:04:22.259000  413772122  119.041985  ...   511     0     NaN
3        2023-06-22 01:04:22.373000  413398520  119.897330  ...   301     0     NaN
4        2023-06-22 01:04:22.402000  413781864  121.010895  ...     0     0     NaN
...                             ...        ...         ...  ...   ...   ...     ...
6467278  2023-06-26 17:13:25.714000  533180188  120.931335  ...   135     0     NaN
6467279  2023-06-26 17:13:25.923000  412630080  120.342200  ...    52     0     NaN
6467280  2023-06-26 17:13:25.982000  412630080  120.342200  ...    52     0     NaN
6467281  2023-06-26 17:13:25.987000  413256369  120.210200  ...   511     0     NaN
6467282  2023-06-26 17:13:25.988000  413307750  120.939950  ...   319     0     NaN

[6467283 rows x 9 columns]
```

Figure 1. The results from data collection

(2) data preprocessing

The data preprocessing workflow is illustrated in Figure 2, encompassing steps such as time format conversion, data sorting, data deduplication, and missing value imputation.

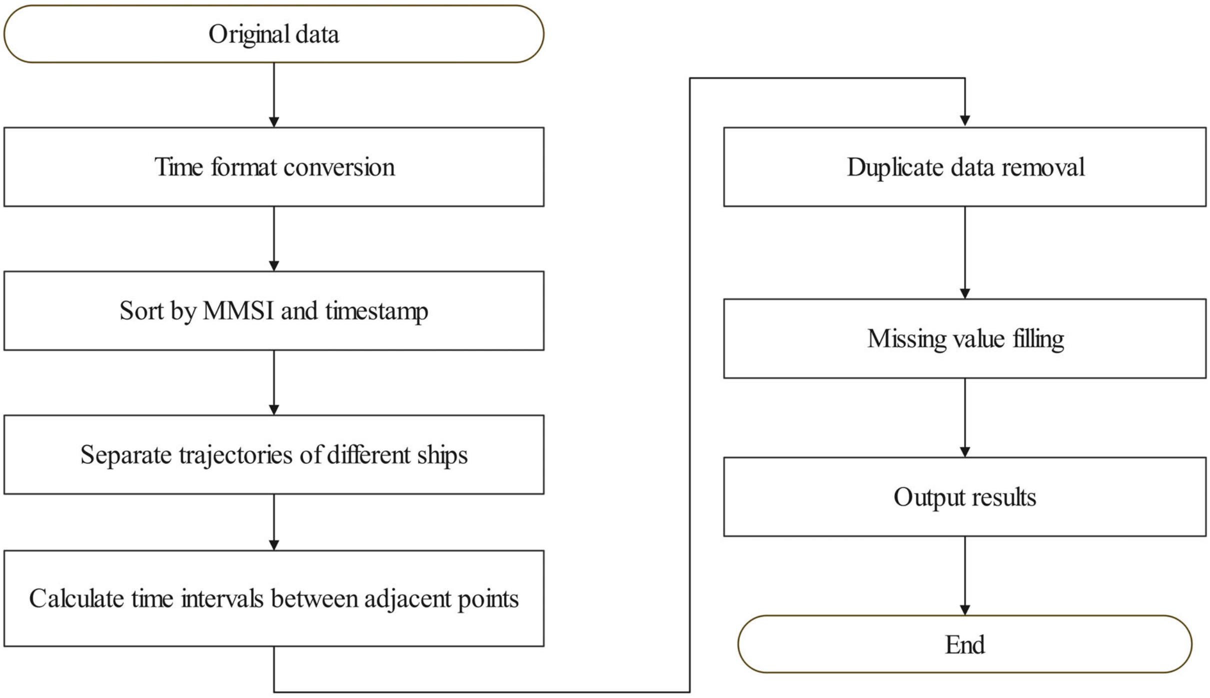

Figure 2. The data preprocessing workflow

Deduplication is a crucial step in data preprocessing, aimed at identifying and removing duplicate data entries within the dataset [7]. In most cases, duplicate data in a dataset may arise from various situations, such as errors during data entry or multiple entries of the same data. Performing deduplication helps enhance the accuracy and consistency of the dataset. Through the deduplication process, the data volume was reduced from 6,467,282 to 6,459,634, with a total of 7,648 duplicate ship navigation records removed.

Missing values, also known as null values, refer to data that is not recorded or unavailable in the dataset [8]. They may occur due to various reasons, such as omissions during data entry or sensor failures. The presence of missing values inevitably affects the results of data analysis,

ISSN: 2710-0588

necessitating missing value imputation before analysis. Different methods, such as mean imputation and backward fill, are employed for different fields.

3.2. Anomaly Detection in Ship Navigation Data

(1) abnormal MMSI detection

In AIS, the nine-digit code typically refers to a part of the MMSI number or a simplified representation used in specific applications. MMSI is a unique identification code used to identify ships and shore-based facilities in maritime radio communications, facilitating communication, tracking, data exchange, and other related operations [9]. Faults in AIS equipment itself, such as damaged antennas or lines, and malfunctions in the main device, can result in the MMSI signal being unable to transmit or display correctly. When ships change their port of registry, name, owner, type, emergency position-indicating radio beacon equipment, or AIS system without first deregistering their MMSI signal and reapplying with the maritime authorities, it can lead to outdated or incorrect MMSI signal information. Conducting anomaly detection on MMSI is also a crucial part of the process. Figure 3 shows the navigation data of some ships with abnormal MMSI.


```
receive,mmsi,lng,lat,course,speed,head,rate,mmsi_str,mmsi_length,mmsi_anomaly,time_diff,trajectory_
466129,2023-06-22 08:48:30.208,0,118,558105,31,909107,32 . 3,4,9,32,0,0,1,abnormal MMSI,0. 0,1,1
898828,2023-06-22 16:20:34.990,0,111,84811,0,0,0,0,0,0,0,0,1,abnormal MMSI,27124,782,2,2
1737422,2023-06-23 08:12:47.078,0,119. 92315,31. 97273,0. 0,0 . 0,0,0,0,1,abnormal MMSI,57132.088,3,
1744953,2023-06-23 08:22:13.404,0,119. 47129,32. 36453,0. 0,0 . 0,0,0,0,1,abnormal MMSI,566 . 326,3,
1767872,2023-06-23 08:52:04.739,0,119,47125,32,364506,0,0,0,0,0,0,0,1,abnormal MMSI,179 1. 335,4,4
1773503,2023-06-23 08:59:26.774,0,119,471214,32,36422,0,0,0,0,0,0,0,1,abnormal MMSI,442,035,4,4
1782627,2023-06-23 09:11:25.130,0,119. 471214,32. 36砬27,0. 0,0. 0,0,0,0,1,abnormal MMSI,718.356,5,
1787467,2023-06-23 09:17:21.719,0,119. 471214,32. 364212,0 . 0,0. 0,0,0,0,1,abnormal MMSI,356. 589,
1796809,2023-06-23 09:30:27.538,0,119,4712,32,36423,0,0,0,0,0,0,0,1,abnormal MMSI,785,819,6,6
1892539,2023-06-23 12:22:08.326,0,119,47128,32,36451,0,0,0,0,0,0,0,1,abnormal MMSI,10300,788,7,7
1922988,2023-06-23 13:00:28.434,0,119. 92315,31. 972713,0,0,0 . 0,0,0,0,1,abnormal MMSI,2300.108,8,
2002074,2023-06-23 14:33:02.438,0,111,84811,0 . 0,0 . 0,0,0,0,0,0,1,abnormal MMSI,5554.004,9,9
2931658,2023-06-24 04:56:55.975,0,111,84811,0,0,0,0,0,0,0,0,1,abnormal MMSI,51833,537,10,10
3047291,2023-06-24 06:54:27.930,0,119,066505,32,2302,220,8,5.1,220,0,0,1,abnormal MMSI,705 1. 955,1
3050841,2023-06-24 06:57:57.896,0,119. 06352,32. 226036,205 . 9,5. 2,205,0,0,1,abnormal MMSI,209.96
3073031,2023-06-24 07:20:14.547,0,111,84811,0 . 0,0,0,0,0,0,0,0,1,abnormal MMSI,1336,651,12,12
3091355,2023-06-24 07:38:34.548,0,111,84811,0,0,0,0,0,0,0,0,1,abnormal MMSI,1100,001,13,13
3126201,2023-06-24 08:14:37.323,0,118,96256,32.174904,272. 3,4,8,272,0,0,1,abnormal MMSI,2162.775,1
3132765,2023-06-24 08:21:37.274,0,118. 95156,32.17507,271. 9,4. 7,271,0,0,1,abnormal MMSI,419.951,1
3136195,2023-06-24 08:25:07.289,0,118,946266,32,175945,285,2,4,8,285,0,0,1,abnormal MMSI,210,015,14
```


Figure 3. Partial Abnormal MMSI

The primary purpose of conducting nine-digit code anomaly detection on MMSI in AIS is to enhance the safety and monitoring efficiency of maritime traffic. By performing nine-digit code anomaly detection, it can be ensured that each ship's MMSI code is unique. This prevents identity confusion caused by incorrect or duplicated MMSI codes, thereby ensuring the uniqueness and accuracy of ship identities.

(2) position and course anomaly detection

The AIS system broadcasts dynamic information such as ships' latitude and longitude, and course, enabling other ships to grasp the real-time dynamics of surrounding ships, thereby effectively avoiding collision accidents [10]. Detecting the latitude and longitude data in AIS can ensure the accuracy and reliability of this information, thereby safeguarding the safety of ship navigation. In some cases, ships may experience abnormal latitude and longitude data due to equipment failures, human operational errors, or other reasons. By detecting the latitude and longitude data in AIS, these anomalies can be promptly identified and addressed. Figure 4 and 5 show the results of latitude and longitude anomaly detection and course anomaly detection, respectively.

ISSN: 2710-0588

```
,receive,mmsi,lng,lat,course,speed,head,rate,lng_anomaly,lat_anomaly,speed_anomaly,course_anomaly
0,2023-06-22 01:04:22.125,413422880,120.84245,31.838917,264.9,3.0,259,0,No,No,False,No
1,2023-06-22 01:04:22.197,413819288,118.96378,32.169735,360.0,0.0,511,-128,No,No,False,Yes
2,2023-06-22 01:04:22.259,413772122,119.041985,32.200928,222.4,4.0,511,0,No,No,False,No
3,2023-06-22 01:04:22.373,413398520,119.89733,32.058674,59.1,0.0,301,0,No,No,False,No
4,2023-06-22 01:04:22.402,413781864,121.010895,31.755903,52.8,0.0,0,0,No,No,False,No
5,2023-06-22 01:04:22.407,413781864,121.010895,31.755903,52.8,0.0,0,0,No,No,False,No
6,2023-06-22 01:04:22.439,413867851,120.93025,31.774353,360.0,0.1,511,-128,No,No,False,Yes
7,2023-06-22 01:04:22.469,412355060,121.02397,31.786356,96.9,4.9,72,0,No,No,False,No
8,2023-06-22 01:04:22.860,248017000,121.06656,31.75785,117.7,13.2,120,7,No,No,False,No
9,2023-06-22 01:04:22.864,248017000,121.06656,31.75785,117.7,13.2,120,7,No,No,False,No
10,2023-06-22 01:04:22.870,413363980,119.05846,32.20725,240.3,0.0,197,0,No,No,False,No
11,2023-06-22 01:04:22.874,413456110,121.03488,31.768578,106.8,6.4,116,-128,No,No,False,No
12,2023-06-22 01:04:22.876,413456110,121.03488,31.768578,106.8,6.4,116,-128,No,No,False,No
13,2023-06-22 01:04:22.879,413456110,121.03488,31.768578,106.8,6.4,116,-128,No,No,False,No
14,2023-06-22 01:04:22.882,413456110,121.03488,31.768578,106.8,6.4,116,-128,No,No,False,No
15,2023-06-22 01:04:22.961,374061000,119.920586,32.119232,329.0,0.1,346,0,No,No,False,No
16,2023-06-22 01:04:22.988,413990949,120.98053,31.76068,3.0,0.0,3,0,No,No,False,No
17,2023-06-22 01:04:22.990,413800182,118.5516,31.9107,211.9,4.0,511,0,No,No,False,No
18,2023-06-22 01:04:22.992,413550570,121.1936,31.670574,130.7,8.5,511,-128,No,No,False,No
19,2023-06-22 01:04:23.015,413550570,121.1936,31.670574,130.7,8.5,511,-128,No,No,False,No
20,2023-06-22 01:04:23.043,412000000,181.0,91.0,360.0,102.3,511,-128,Yes,Yes,True,Yes
21,2023-06-22 01:04:23.092,413829168,118.99094,32.178196,253.5,3.1,253,0,No,No,False,No
22,2023-06-22 01:04:23.159,413342420,120.9198,31.81393,324.9,9.2,333,0,No,No,False,No
```

Figure 4. The results of latitude and longitude anomaly detection

```
,receive,mmsi,lng,lat,course,speed,head,rate,lng_anomaly,lat_anomaly,speed_anomaly,course_anomaly
0,2023-06-22 01:04:22.125,413422880,120.84245,31.838917,264.9,3.0,259,0,No,No,False,No
1,2023-06-22 01:04:22.197,413819288,118.96378,32.169735,360.0,0.0,511,-128,No,No,False,Yes
2,2023-06-22 01:04:22.259,413772122,119.041985,32.200928,222.4,4.0,511,0,No,No,False,No
3,2023-06-22 01:04:22.373,413398520,119.89733,32.058674,59.1,0.0,301,0,No,No,False,No
4,2023-06-22 01:04:22.402,413781864,121.010895,31.755903,52.8,0.0,0,0,No,No,False,No
5,2023-06-22 01:04:22.407,413781864,121.010895,31.755903,52.8,0.0,0,0,No,No,False,No
6,2023-06-22 01:04:22.439,413867851,120.93025,31.774353,360.0,0.1,511,-128,No,No,False,Yes
7,2023-06-22 01:04:22.469,412355060,121.02397,31.786356,96.9,4.9,72,0,No,No,False,No
8,2023-06-22 01:04:22.860,248017000,121.06656,31.75785,117.7,13.2,120,7,No,No,False,No
9,2023-06-22 01:04:22.864,248017000,121.06656,31.75785,117.7,13.2,120,7,No,No,False,No
10,2023-06-22 01:04:22.870,413363980,119.05846,32.20725,240.3,0.0,197,0,No,No,False,No
11,2023-06-22 01:04:22.874,413456110,121.03488,31.768578,106.8,6.4,116,-128,No,No,False,No
12,2023-06-22 01:04:22.876,413456110,121.03488,31.768578,106.8,6.4,116,-128,No,No,False,No
13,2023-06-22 01:04:22.879,413456110,121.03488,31.768578,106.8,6.4,116,-128,No,No,False,No
14,2023-06-22 01:04:22.882,413456110,121.03488,31.768578,106.8,6.4,116,-128,No,No,False,No
15,2023-06-22 01:04:22.961,374061000,119.920586,32.119232,329.0,0.1,346,0,No,No,False,No
16,2023-06-22 01:04:22.988,413990949,120.98053,31.76068,3.0,0.0,3,0,No,No,False,No
```

Figure 5. The results of course anomaly detection

(3) speed anomaly detection

The AIS system provides an effective obstacle avoidance measure for ships by monitoring their speed. When ships are sailing at sea, speed information is a crucial parameter for judging their dynamic trajectory and driving status. By obtaining real-time speed information of ships, the AIS system can more accurately predict the ship's sailing trajectory, thereby preventing collision accidents. For the determination of speed anomalies, this paper adopts the IQR (Interquartile Range) method to flag and handle outlier values, and the result is shown in Figure 6.

ISSN: 2710-0588

```
,receive,mmsi,lng,lat,course,speed,head,rate,lng_anomaly,lat_anomaly,speed_anomaly,course_anomaly
0,2023-06-22 01:04:22.125,413422880,120.84245,31.838917,264.9,3.0,259,0,No,No,False,No
1,2023-06-22 01:04:22.197,413819288,118.96378,32.169735,360.0,0.0,511,-128,No,No,False,Yes
2,2023-06-22 01:04:22.259,413772122,119.041985,32.200928,222.4,4.0,511,0,No,No,False,No
3,2023-06-22 01:04:22.373,413398520,119.89733,32.058674,59.1,0.0,301,0,No,No,False,No
4,2023-06-22 01:04:22.402,413781864,121.010895,31.755903,52.8,0.0,0,0,No,No,False,No
5,2023-06-22 01:04:22.407,413781864,121.010895,31.755903,52.8,0.0,0,0,No,No,False,No
6,2023-06-22 01:04:22.439,413867851,120.93025,31.774353,360.0,0.1,511,-128,No,No,False,Yes
7,2023-06-22 01:04:22.469,412355060,121.02397,31.786356,96.9,4.9,72,0,No,No,False,No
8,2023-06-22 01:04:22.860,248017000,121.06656,31.75785,117.7,13.2,120,7,No,No,False,No
9,2023-06-22 01:04:22.864,248017000,121.06656,31.75785,117.7,13.2,120,7,No,No,False,No
10,2023-06-22 01:04:22.870,413363980,119.05846,32.20725,240.3,0.0,197,0,No,No,False,No
11,2023-06-22 01:04:22.874,413456110,121.03488,31.768578,106.8,6.4,116,-128,No,No,False,No
12,2023-06-22 01:04:22.876,413456110,121.03488,31.768578,106.8,6.4,116,-128,No,No,False,No
13,2023-06-22 01:04:22.879,413456110,121.03488,31.768578,106.8,6.4,116,-128,No,No,False,No
14,2023-06-22 01:04:22.882,413456110,121.03488,31.768578,106.8,6.4,116,-128,No,No,False,No
15,2023-06-22 01:04:22.961,374061000,119.920586,32.119232,329.0,0.1,346,0,No,No,False,No
16,2023-06-22 01:04:22.988,413990949,120.98053,31.76068,3.0,0.0,3,0,No,No,False,No
17,2023-06-22 01:04:22.990,413800182,118.5516,31.9107,211.9,4.0,511,0,No,No,False,No
18,2023-06-22 01:04:22.992,413550570,121.1936,31.670574,130.7,8.5,511,-128,No,No,False,No
19,2023-06-22 01:04:23.015,413550570,121.1936,31.670574,130.7,8.5,511,-128,No,No,False,No
20,2023-06-22 01:04:23.043,412000000,181.0,91.0,360.0,102.3,511,-128,Yes,Yes,True,Yes
21,2023-06-22 01:04:23.092,413829168,118.99094,32.178196,253.5,3.1,253,0,No,No,False,No
```

Figure 6. The results of speed anomaly detection

3.3. Anomaly Detection of Ship Navigation Trajectories

K-means algorithm is a widely used clustering analysis algorithm that partitions the samples in a dataset into K non-overlapping clusters through an iterative solution. The samples within each cluster are as close as possible to the cluster center, while the cluster centers of different clusters are as far apart as possible. The K-means algorithm requires the user to specify the number of categories in the data sample, as the number of clusters needs to be determined using the Elbow method before clustering the data, which is then combined with K-means for data processing. K-means is an unsupervised learning-based clustering method, commonly used for outlier detection [11], and the cluster result is stored in the 'AnomalyDetectedDataB.csv' file. The cluster result is shown in Figure 7.

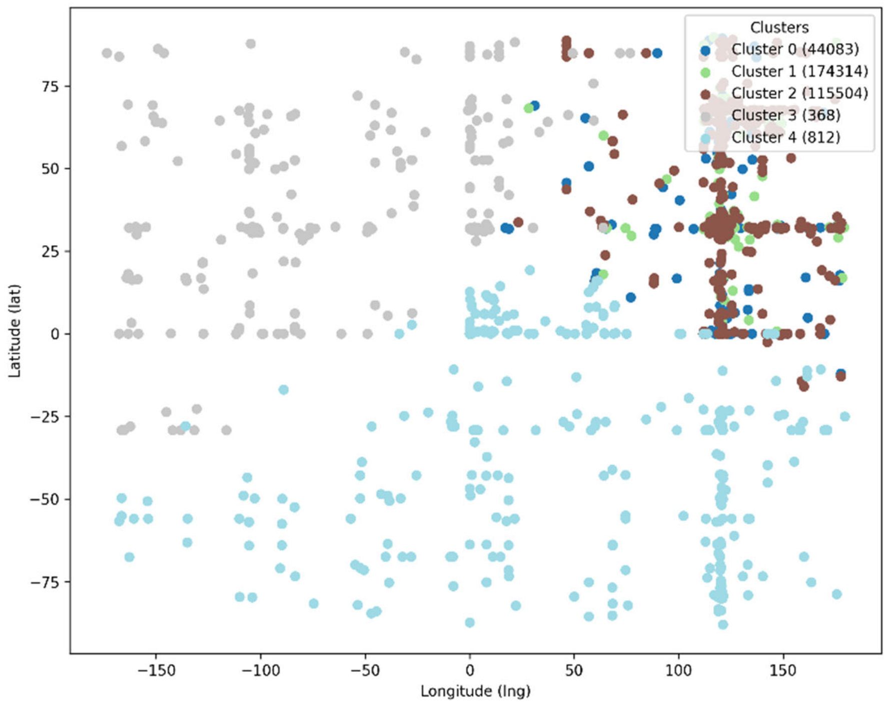

Figure 7. The clustering of ship navigation data

ISSN: 2710-0588

4. Conclusion

This paper designs an anomaly detection algorithm with high accuracy and reliability. Experimental results demonstrate that this algorithm can accurately identify abnormal behaviors of ships and issue timely warnings, providing powerful decision support for shipping management departments. Additionally, the algorithm exhibits good versatility and scalability, enabling it to adapt to different types and scales of ship data, thus possessing broad application prospects.

Data quality is one of the key factors influencing the effectiveness of ship anomaly behavior detection. In future research, it is necessary to strengthen efforts in the collection, processing, and storage of AIS data to enhance data quality and standardization. Simultaneously, establishing universal data standards and sharing mechanisms is essential to facilitate data exchange and sharing among different systems. The technology for anomaly behavior detection also needs to be optimized to achieve real-time monitoring of abnormal behaviors and provide more precise and immediate feedback for maritime traffic safety. Looking ahead, the detection of ship anomaly behaviors based on data mining should place greater emphasis on technological integration and innovation. Combining advanced technologies such as reinforcement learning can improve the accuracy and efficiency of ship anomaly behavior recognition. Furthermore, integrating ship anomaly behavior detection with other technologies, like the Internet of Things, can enable more comprehensive and intelligent shipping safety management.

Acknowledgements

This work was financially supported by the funding of the Research Program of Jiangsu Maritime Institute(2022ZKyb02), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (22KJB580002), and the Excellent Teaching Team for QingLan Project of the Jiangsu Higher Education Institutions of China (Big Data Technology Teaching Team with Shipping Characteristic).

References

[1] Rong, Hao, A. P. Teixeira, and C. Guedes Soares. "Data mining approach to shipping route characterization and anomaly detection based on AIS data." Ocean Engineering 198 (2020): 106936.

[2] Filipiak, Dominik, et al. "Big data for anomaly detection in maritime surveillance: spatial AIS data analysis for tankers." Maritime Technical Journal 215.4 (2018): 5-28.

[3] Liu, Hongdan, et al. "Research on detection mechanism of vessel abnormal behavior based on immune genetic spectral clustering algorithm." Ocean Engineering 266 (2022): 113099.

[4] Ikotun, Abiodun M., et al. "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data." Information Sciences 622 (2023): 178-210.

[5] El Bouchefry, Khadija, and Rafael S. de Souza. "Learning in big data: Introduction to machine learning." Knowledge discovery in big data from astronomy and earth observation. Elsevier, 2020. 225-249.

[6] Landi, Federico, et al. "Working memory connections for LSTM." Neural Networks 144 (2021): 334-341.

[7] Akbar, Mohd, et al. "Enhanced authentication for de-duplication of big data on cloud storage system using machine learning approach." Cluster Computing 27.3 (2024): 3683-3702.

[8] Emmanuel, Tlamelo, et al. "A survey on missing data in machine learning." Journal of Big data 8 (2021): 1-37.

ISSN: 2710-0588

[9] Guan, Meng, Yu Cao, and Xu Cheng. "Research on the recognition of multiple MMSI codes on a single vessel based on AIS datas." IEEE Access (2023).

[10] Rong, H., A. P. Teixeira, and C. Guedes Soares. "Ship collision avoidance behaviour recognition and analysis based on AIS data." Ocean Engineering 245 (2022): 110479.

[11] Ahmed, Mohiuddin, Raihan Seraj, and Syed Mohammed Shamsul Islam. "The k-means algorithm: A comprehensive survey and performance evaluation." Electronics 9.8 (2020): 1295.