# Vessel track information mining using AIS data

Feng Deng*‡, Sitong Guo*, Yong Deng*, Hanyue Chu*, Qingmeng Zhu* and Fuchun Sun†
*Science and Technology on Integrated Information System Laboratory
Institute of Software, Chinese Academy of Sciences, Beijing, China
Email: guositong2006@126.com
†Department of Computer Science and Technology
Tsing Hua University, Beijing, China
‡ University of Chinese Academy of Sciences, Beijing, China

***Abstract*—In recent years, vessel traffic and maritime situation awareness become more and more important for countries across the world. *AIS data* contains much information about vessel motion and reflects traffic characteristics. In this paper, *data mining* is introduced to discover motion patterns of vessel movements. Firstly, we do statistical analysis for large scale of AIS data. Secondly, we use *association rules* to analyze the frequent moving status of vessels. We extend the dimensions of data features, improve the algorithm in efficiency and import the concept of time scale in the algorithm based on the previous relative work. Thirdly, we introduce *Markov model* to make supplement for the association rules. The prediction results in the Markov process are further used to do the anomaly detection. The method in this paper provides novel idea for the research in AIS data and the management of maritime traffic.**

***Keywords*—*AIS data, data mining, Association rules, Markov***

## I. Introduction

As economy develops, the enhancement of waterway transport and International Trade leads to the rise of maritime traffic. Considering that more than 90% of international trading about China is carried by sea, the transport analysis and surveillance of maritime for China has great importance. Traditionally, tracing the vessel and detecting the maritime situation are through equipment as radar, infrared etc. Nowadays with the development of technology, the self reporting system has been introduced to trace the vessel and surveil the maritime. The most widely applied self-reporting system is Automatic Identification System, which has been introduced by International Maritime Organization and International Convention for the Safety of Life at Sea (SOLAS). Most of civilian vessels such as cargo ships, container ships, tugs and fishing vessels are required compulsory to be equipped with AIS facility. With the widely application of Automatic Identification System, large AIS data are generated. Through analyzing the attributes of the AIS data such as latitude, longitude, we can mine the vessel motion patterns, predict the motion status and make detection of anomaly motion status.

Several methods have been proposed in mining AIS data and analyzing vessel motions. Feixiang Z[1] has used association rules to discover the association positions of the vessel, and gives the example of detecting anomaly through his method. But he does not consider the time scale of tracks. Nicholson AE[2] models the data by Bayesian network. Supplementary data are introduced to get more characteristics. And she evaluates the anomaly in both static and dynamic method. The wealth of the complementary attributes made her method reliable and hard to override. In some papers, statistical related methods are introduced. B.Ristic[3] carries out motion anomaly detectors in the framework of adaptive kernel density estimation. And motions of vessels are predicted using Gaussian sum tracking filter according to the history data. Methods of cluster are introduced as well. Kraiman[4] clusters the normal motion by Gaussian Mixture Model, and makes anomaly detection through the result. Dahlbom[5] clusters the track of the vessels to find out the isolated point. And the methods of mining visualization have been introduced. Falkman[6] emphasizes the import of domain knowledge, which is an interactive way to analyze the motion of vessels. Other methods such as Hidden Markov have been used. Tun[7] uses Hidden Markov to judge the classes the vessel belongs to. In this way, he distinguishes normal and anomaly in a port. But he considers only the position sequence of a vessel, while the speed and course over ground are also important in analysis of vessel motion.

The paper is structured as follows. In the first session of this paper, we make a brief introduction of the backgrounds and related work. In the second session, we introduce the methods for data preprocessing. In the third session, we do statistical wok for the data, and cluster briefly to show the data that to be processed. In the fourth session, the algorithm of Fp-Growth for association rules is introduced to learn the frequent status of vessels in the data set. In the fifth session, we model the data with Markov chain, and introduce the matrix of transaction probability for vessel statuses. In the sixth session, we analyze the results and give an example of anomaly detection. In the seventh session of the paper, we make a conclusion of our work.

## II. Data Preprocessing

In this session, we do data preprocessing, including data selecting, data cleaning and data discretizing.

### A. Data Selecting

We choose limited area and time period from raw data. We use AIS data from $11.1^{st}$ 2013 to $11.30^{th}$ 2013, which distributes from longitude E 119.833° to E 122.667°, latitude N 30.75° to N 32.417°. There are over 2 million records in the data set. And we separate it into 3 partitions, one training set and two testing sets. Every partition is continues in time stamp.

### B. Data Cleaning

Firstly, we delete data without MMSI, position, speed, course over ground or time stamp. Secondly, we fill the records without ship name, ship type which can be gotten from other records with the same MMSI. Thirdly, we drop the data with same time stamp and same location. Fourthly, we wipe off some attributes that are not to be used such as year built. The final attributes left and the relative explanations are shown in table I.

TABLE I. THE LEFT ATTRIBUTES WITH EXPLANATIONS

| Attributes | Explanations |
|---|---|
| MMSI | Maritime Mobile Service Identify of ships, nine digital with type of long int |
| SHIPTYPE | Ship types, words as "Cargo" with type of string |
| SHIPNAME | Ship names, words as "Moon" with type of string |
| FLAG | Ship flags, words as "China" with type of string |
| COURSE | Course over ground of ships, numeric as "105.5°" with type of float (North is 0°) |
| LONGTITUDE | Longitude of ships, numeric as "119.5°" with type of float |
| LATITUDE | Latitude of ships, numeric as "32.5°" with type of float |
| SPEED | Speed of ships, numeric as "15.5" with type of float (km/h) |
| POSITION TIME | Position time of the record as "2013/11/14 12:23:24" with type of string |

After cleaning the raw data, we sort the result by MMSI firstly and POSITION TIME secondly. This process aims at discovering the sing-track of each ship. The following algorithm of association rules cannot guarantee the time scales for each record. But after find the one-direction for every single-track of each vessel, we can learn the dependence of the motion status according to their locations. Further work such as adding the time stamp for the records will be conducted in the fourth session.

### C. Data discretizing

In order to suppress the computation complexity and make the data more appropriate for the model in the following session. We discretize attributes of AIS records into different statuses, especially for continuous position information. We use grids of 0.1° × 0.1° to dived up the area (as shown in Fig.1), and the AIS position is allocated into the grid. For example, "3_11" means that the position of this ship is in row 3 and column 11. The discretion method of the data is inherited from Feixiang Z[1].

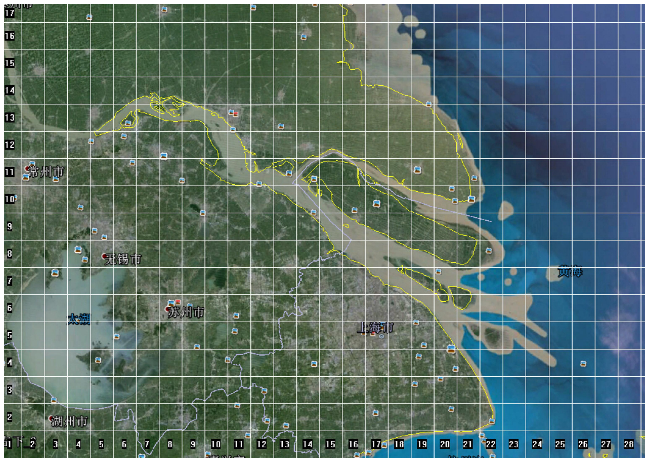

Fig. 1. Position discrete result.

According to the frequency standard of AIS dynamic information[8], we divide the speed into four statuses. Just as shown in table II.

TABLE II. DISCRETE OF SPEED

| Range of Speed (Km/h) | Discrete values |
|---|---|
| 0 ∼ 3 | Slow |
| 3 ∼ 14 | Medium |
| 14 ∼ 23 | High |
| 23 ∼ 99 | Very High |
| Over 99 | Exception |

After discretizing the speed of the data, we can find the anomaly data, which is labeled "Exception". The appearance of the "Exception" is sudden and discontinuous, which is probably the reason of data fault. In this paper, we do not input records with "Exception" into the following algorithm and model.

According to the course information of the data with the knowledge about the weather over sea, we divide the course over ground into 8 statuses. And the interval of each value is 45°, as shown in table III.

TABLE III. DISCRETE OF COURSE OVER GROUND

| Course over ground | Discrete values |
|---|---|
| 337.5° ∼ 22.5° | N |
| 22.5° ∼ 67.5° | NE |
| 67.5° ∼ 112.5° | E |
| 112.5° ∼ 157.5° | SE |
| 157.5° ∼ 202.5° | S |
| 202.5° ∼ 247.5° | SW |
| 247.5° ∼ 292.5° | W |
| 292.5° ∼ 337.5° | NW |

After data preprocessing, there are about 60% of the raw data left. But they will not be the direct input data into the fourth and fifth session. Final input data are decided by statistics and cluster results in the third session.

## III. STATISTIC AND CLUSTER

In this session, we do simple data statistics and data clusters.

### A. Statistics

Statistics of the data focuses on the type of the vessels, the flag of the vessels, the speed of the vessels and the course over ground of the vessels in the data set that has been gotten in the second session. The result is shown in Fig.2.

As shown in Fig.2 vessels with type "Cargo" and flag "China" possess the most percentage in the total data. The speed of the vessel is mostly "Slow" and "Medium". Considering the area we choose is near Shang Hai port, which means the density of the vessel in the area is very high, the result that vessels do not move fast is in accordance with the reality.

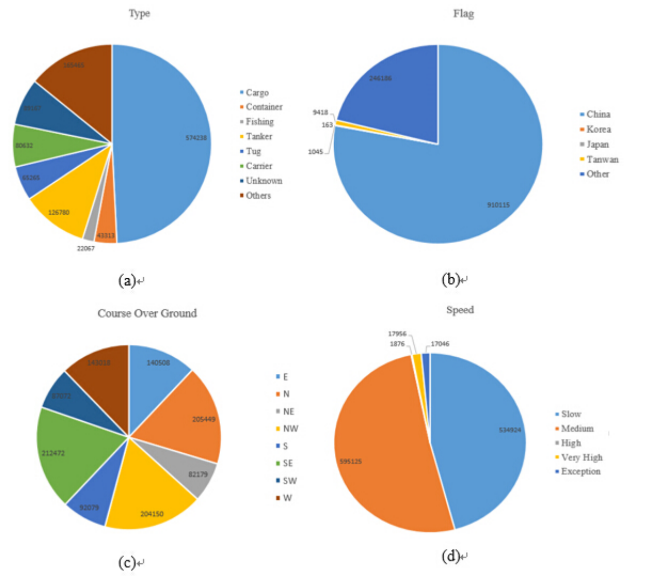

Fig. 2. Statistical result (a) type (b) Flag (c) Speed (d) Course over ground.

### B. Cluster

According to the result of data statistics, we cluster the data by vessel flag and vessel type. And we can view the data with different type and flag "China" in Fig.3.

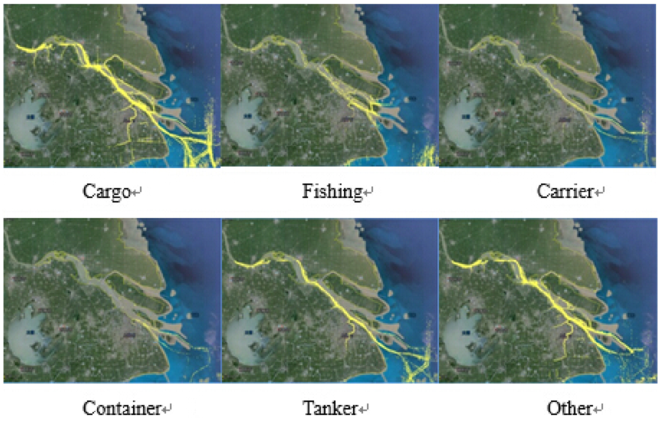

Fig. 3. Clusters.

We witness that the type "Cargo ship" and type "Other" with flag "China" presents clear and relative focused track compared with type fish, carrier and so on as shown in Fig.3. And the Cargo ship with flag "China" possess relative larger records than the type "Other" as shown in Fig.2(a).

In this way, the data with the type "Cargo" and flag "China" are chosen. And they will be the input data in the following analysis.

## IV. ASSOCIATION RULES

In this session, we adopt the theory about association rules to discover the relationships of frequent statuses of vessels.

Association rules have been introduced in 1993[9], which are used to verify the relationships of items in a data set. Both Apriori algorithm and Fp-Growth algorithm describes the theory. But Fp-Growth algorithm is more suppressed in the computation complex. This paper will apply Fp-Growth algorithm in the mining of AIS data with the import of time scale.

In the previous work about association rules in AIS data, the only dimension is the discrete location. In this paper we extern the dimensions of the input data, for example "6_18_E_Medium" represents the status that the vessel location is 6_18 (rectangle in row 6 column 18 shown in Fig.1), with course in the range of 67.5° $\sim$ 112.5° (as in table III), and speed in the range of 3$\sim$14 (as in table II).

In this paper the time stamp of the specific status is introduced during the calculation of Fp-Growth. We divided time period every one hour. And attach the time period in the vessel status. For example, the "6_18_E_Medium_01_02" means that this status "6_18_E_Medium" is witnessed during 01:00 and 02:00. Thus the orders of statuses can be confirmed. And we split the training data into different data sets according to the process the vessels move. Every set of the data store one process of the vessels. In this way, the order of status sequences of one vessel can be clearly understood according to the attached time stamp and the location, for the raw data have already been sorted according to the process of movement.

Then we reflect the statuses with different IDs. Some examples of the reflection relationships are shown in table IV.

TABLE IV. Reflect From Statuses To Ids

| Status | Reflection IDs |
|---|---|
| 10_1_E_Medium_01_02 | 1 |
| 10_10_NE_Hign_06_07 | 2 |
| 12_4_NW_Low_10_11 | 3 |
| 13_3_SE_VeryHigh_15_16 | 4 |
| 15_17_E_Low_21_22 | 5 |

After the reflection from statuses to ids, we can get the id sequences which represent the statuses that vessels move. The style it presents is the same with the classic Market Basket Analysis (as shown in table V).

TABLE V. Input data for Market Basket (TID) and Vessels Statuses Sequences (MMSI)

| TID | ITEMS | MMSI | Reflection IDs |
|---|---|---|---|
| T1 | 1,4,5 | 1XXXX | 1,4,5,9,8 |
| T2 | 2,6,9 | 2XXXX | 2,6,8,1,5 |
| T3 | 3,3,4 | 3XXXX | 3,3,4,5,7 |
| T4 | 4,1 | 4XXXX | 4,1 |
| T5 | 5 | 5XXXX | 5,8 |

With the data have been processed, two important indicators can be defined.

***Definition 1:*** The support of the association rule is the percentage (ratio) of records to be processed in the total records of the data set.

***Definition 2:*** The confidence of association rule $X \rightarrow Y$ is the probability of the records containing $Y$ which have been already containing $X$.

In this paper the support of the association rule is defined 0.01. Compared with Apropri algorithm, Fp-Growth does not

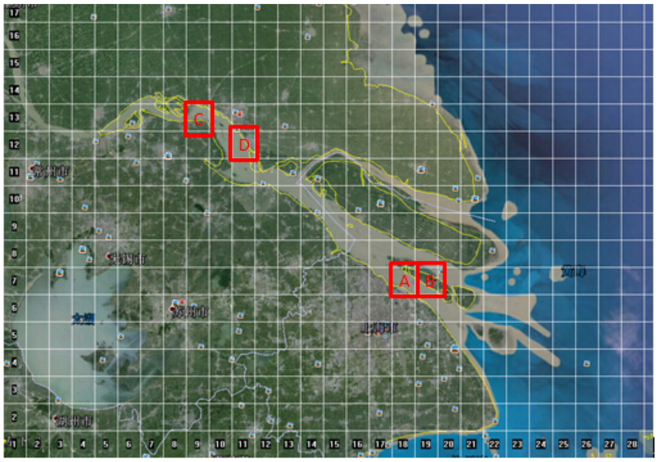

Fig. 4. Result present of Association rules.

produce candidates for frequent items during the processing period. For this reason, it has great advantage in time and space complexity.

The result of the algorithm is the association relationships between different statuses with time stamp. Examples are shown in table VI.

TABLE VI. Example Of Association Rules With Confidence

| ID | Status | ID | Status | Confidence |
|---|---|---|---|---|
| 1796 | 7_18_SE_Medium_05_06 | 1787 | 7_19_E_Slow_08_09 | 0.92 |
| 1824 | 13_9_E_Slow_06_07 | 1891 | 12_11_SE_Medium | 0.89 |
| ...... | ...... | ...... | ...... | ...... |

The rules of 1796→1787 means that if we witness a vessel in position 7_18 (shown in Fig.4 A) with medium speed (3 $\sim$ 14) and directing SE (112.5° $\sim$ 157.5°), the following status will be in position 7_18 (shown in Fig.4 B) with low speed (0 $\sim$ 3) and directing E (67.5° $\sim$ 112.5°).

But the result of association rule is uncertain in continuous. For example, we get “13_9_E_Slow→12_11_SE_Medium”, the position of 13_9 and 12_11 is not neighbored (as shown in Fig.4 C and D). To analyze the progress between statuses “13_9_E_Slow” and “12_11_SE_Medium”, we introduce the method of Markov in the fifth section.

## V. Markov Model

In this session, we model the AIS data with Markov. Then make a general introduction to its prediction and anomaly detection.

Markov model describes a transition process from one status to another, with the property that next status depends only on the current one. The movements of vessels have the same feature. The next status of a vessel, which means the location it will be , the speed it will has, and the course it will direct, is only related to the current location it is, the current speed it has and the current course over ground it directs.

***Definition 1:*** Let random variables $X_0$, $X_1$, $X_2$ represent status sequence. The possible values of $X_i$ form a countable set S called the status space of Markov. Let the $X_n$ represents the status in time n. If the sequence has the property:

$$\mathrm{P}(X_{n+1} = \mathrm{x} \mid X_0, X_1, X_2, \ldots, X_n) = \mathrm{P}(X_{n+1} = \mathrm{x} \mid X_n)$$

Then the variables form a Markov chain and this property is the Markov chain property. The model that Markov chain belongs is Markov model.

***Definition 2:*** Let *P* be the matrix of transition probability. Let $P_{ij}$ represent the transition probability from status $i$ to status $j$ . Then the probability of status $i$ to status $j$ after 2 steps is:

$$P_{ij}^{(2)} = \sum_{k=1}^{r} P_{ik}P_{jk}$$

In this way, the probability $P_{ij}^{n}$ which means status $i$ to status $j$ after $n$ steps can be calculated.

Markov model is regularly used in the field of natural language processing. The model makes prediction for next letter or word according to the history training set. The application of Markov model in position statuses and tracks is usually for GPS data[10][11]. For AIS data, only Tun applied hidden Markov model in the research about port management. And he classifies vessel tracks to different classes through the model. The tracks without any known classes are labeled as dangerous. In this way the surveillance of port is strengthened.

In this paper, we use Markov in another way, the transaction probability matrix (as in table VII) is learned from training data. The transaction of status routines (as in Fig.5) is constructed based on the matrix of transition probability.

TABLE VII. Example Of Transaction Matrix

| Status | 13_9_E_<br>Slow | 13_10_E_<br>Slow | 12_10_SE_<br>Medium | 12_11_SE_<br>Medium |
|---|---|---|---|---|
| 13_9_E_Slow | 0.133 | 0.583 | 0.280 | 0 |
| 13_10_E_Slow | 0 | 0.215 | 0.418 | 0.367 |
| 12_10_SE_Medium | 0 | 0 | 0.13 | 0.57 |
| 12_11_SE_Medium | 0 | 0 | 0 | 0.274 |

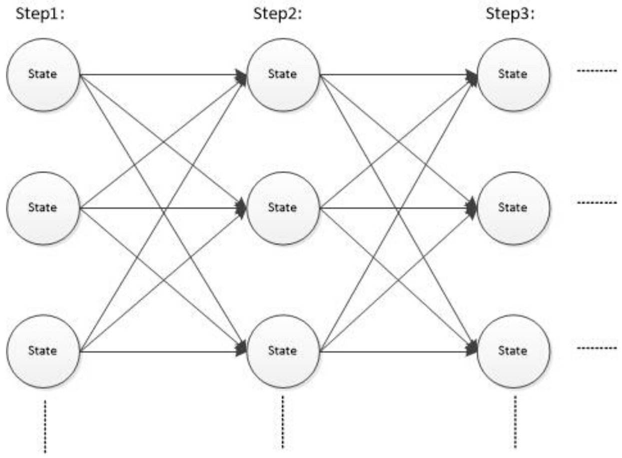

Fig. 5. Transaction Status Routines.

From the transaction statuses, we can predict the next status for a specific vessel based on its current status. And we can predict the next $n$ status of this vessel in this way. We can describe the routine of this vessel with the probability for very steps. In this paper we mainly use the results of one step transaction probability of vessels to make predictions and anomaly detections.

For example, if the current status of the specific vessel is “13_10_E_Slow”. According to the transaction matrix in table VII, the most probability of next status is “12_10_SE_Medium” with the probability 0.418, and the other two probable statues are “13_10_E_Slow” and “12_11_SE_Medium” which have the probability 0.215 and 0.367 separately.

In this way we make predictions of the next statues. If next status is not the three statues listed above, in other words, if the next status is the one with probability 0, we mark it anomaly. For the multifarious dimensions of the input data, we can easily get the specific anomaly reason. For example, if the next status of the specific vessel is “12_10_NW_Medium”, we compare it with the normal status “12_10_SE_Medium” and can get the reason for anomaly is the course over ground “NW” easily.

## VI. Result Analysis

In this session, we introduce how the result of Markov model can make supplement of association rules. And we give an example for anomaly detection using the results.

In this paper, association rules give the hot location and statuses while Markov model gives continuous status transaction probabilities. The combination of association rules and Markov model can provide a relative comprehensive surveillance of the vessel motion and maritime situation. For example, from the dependence relationships in table VI we know the key positions are 13_9 (shown with C in Fig.4) and 12_11 (shown with D in Fig.4). So the vessel in 13_9 with status “13_9_E_Slow” needs the highest attention. For its track statuses have been learned in the training set. Here we give an example of a vessel (shown in table VIII) that can be traced through this method.

One vessel (information in table VIII) has been witnessed in the hot location 13_9 with status “13_9_E_Slow” And now we trace its statuses. According to the training set and matrix in table VII, we simulate its routine (as shown in Fig.6). And the analysis progress of its movement is shown in table IX.

TABLE VIII. Vessel Information

| | |
|---|---|
| **MMSI** | 4XXXX |
| **FLAG** | China |
| **SHIPNAME** | XXXX |
| **SHIPTYPE** | Cargo Ship |
| **LATITUDE** | 3XXX |
| **LONGITUDE** | 1XXX |
| **SPEED** | X Km/h |
| **COURSE OVER GROUND** | $X°$ |
| **STATUS** | 13_9_E_Slow |

In table IX, every motion of the vessel is calculated. The probability for the current status means that with the certain precursor status how possible it can change to the current one. And we calculate every probability of next status for this vessel. In this way, continuous movement of the vessel is traced.

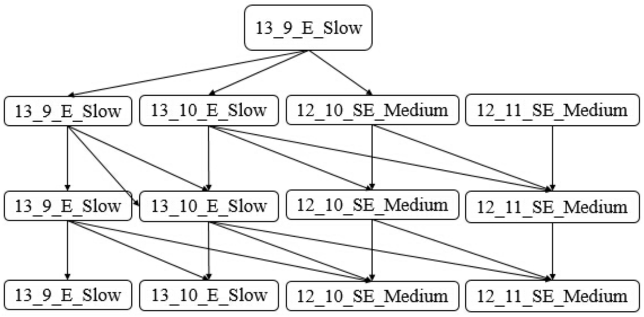

Fig. 6. Transaction Status Routines of vessel 4XXXX.

TABLE IX. Continuous Movement Of 4XXXX

| MMSI | Position Time | Status | Probability |
|---|---|---|---|
| 4XXXX | 2X/X/X X:X | 13_9_E_Slow | Start |
| 4XXXX | 2X/X/X X:X | 13_10_E_Slow | 0.583 |
| 4XXXX | 2X/X/X X:X | 12_10_SE_Medium | 0.418 |
| 4XXXX | 2X/X/X X:X | 12_11_SE_Medium | 0.57 |

In Fig.6, we get the statuses transaction routine. According to the routine and the matrix in table VII, we can get the probability of statues in every step. We assume that the starting status is “13_9_E_Slow”, and then the following n steps can be calculated (as show in table X with 3 steps).

TABLE X. Example Of Transaction Matrix In Every Steps

| Status | 13_9_E_ Slow | 13_10_E_ Slow | 12_10_SE_ Medium | 12_11_SE_ Medium |
|---|---|---|---|---|
| Step 1 | 0.133 | 0.583 | 0.280 | 0 |
| Step 2 | 0 | 0.215 | 0.418 | 0.367 |
| Step 3 | 0 | 0 | 0.13 | 0.57 |

The probability in this routine from start to the status “12_10_SE_Medium” in step 2 and step 3 is the highest. This result is in accordance with the relationship learned from association rules.

In the second session, we divided the test set into 2 parts. In the first test set, we learned the minimum probability $P_{min}$. The probability serves as threshold. In the second test set, the prediction probability that is larger than the threshold is classified normal, while the lower classified anomaly. In this paper, the $P_{min}$ is 0.107.

In this way, the anomaly detection is processed (as shown in table XI).

TABLE XI. Example Of Anomaly Detection

| MMSI | Status | Probability | Classification |
|---|---|---|---|
| 3XXXX | 8_16_SE_Medium | 0.208 | Normal |
| 3XXXX | 8_17_SE_Medium | 0.488 | Normal |
| 3XXXX | 7_17_SE_Medium | 0.31 | Normal |
| 3XXXX | 7_18_SE_Medium | 0.515 | Normal |

## VII. Conclusion

In this paper, modified association rules and Markov model are introduced in data mining for AIS data. The dimensions of characteristics are increased. So the results contains much more information for analysis. In the process of FP-Growth, the import of the time scale concept makes the results more exact and more reliable. And in the following Markov model, continuous statuses transaction makes sufficient supplement for the results of association rules. The method mentioned in this paper can be applied in the management of waterways and surveillance of maritime situation.

The future work will be focused on the improving of the result. In the data preprocessing period, fusing data in a fix time will suppress the negative impact of the data that not continuous in time. In the cluster period, clustering the data according to tracks will make the association rules more targeted. Other algorithms and methods of data mining and machine learning will be considered in the application of AIS data as well.

## ACKNOWLEDGEMENT

The authors are grateful to the research supervisor, who gives courage, confidence, and guidance during the research work. And the colleagues who help carry out the experiment and collect data.

And This work was supported by Beijing Municipal Science Foundation No. 4133092 and 863 program (No.2012AA011206).

## REFERENCES

[1] Z Feixiang. “Mining ship spatial trajectory patterns from AIS database for maritime surveillance,” Emergency Management and Management Sciences (ICEMMS), 2011 2nd IEEE International Conference on. IEEE, 2011, pp. 772–775.

[2] S. Mascaro, A. E. Nicholson, K. Korb. “Anomaly detection in vessel tracks using bayesian networks,” International Journal of Approximate Reasoning, 2013.

[3] B. Ristic, B. La Scala, M. Morelande et al. “Statistical analysis of motion patterns in AIS data: Anomaly detection and motion prediction,” Information Fusion, 2008 11th International Conference on. IEEE, 2008, pp. 1–7.

[4] J. B. Kraiman, S. L. Arouh, M. L. Webb. “Automated anomaly detection processor,” AeroSense 2002. International Society for Optics and Photonics, 2002, pp. 128–137.

[5] A. Dahlbom, L. Niklasson. “Trajectory clustering for coastal Surveillance,” Information Fusion, 2007 10th International Conference on. IEEE, 2007, pp. 1–8.

[6] M. Riveiro, G. Falkman. “The role of visualization and interaction in maritime anomaly detection,” IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, 2011, pp. 78680M-78680M-12.

[7] M. H. Tun, G. S. Chambers, T. Tan et al. “Maritime port intelligence using AIS data,” Recent advances in security technology, 2007, pp. 33.

[8] I. Recomendations. 1371-1, “Technical characteristics for universal shipborne automatic identification system using time division multiple access in the VHF maritime mobile band,” 2001.

[9] R. Agrawal, T. Imieliski, A. Swami. “Mining association rules between sets of items in large databases,” ACM SIGMOD Record. ACM, 1993, vol. 22(2), pp. 207–216.

[10] A. Witayangkurn, T. Horanont, Y. Sekimoto et al. “Anomalous event detection on large-scale GPS data from mobile phones using hidden markov model and cloud platform,” Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication, ACM, 2013, pp. 1219–1228.

[11] D. Ashbrook, T. Starner. “Learning significant locations and predicting user movement with GPS,” Wearable Computers, 2002(ISWC 2002). Proceedings. Sixth International Symposium on. IEEE, 2002, pp. 101–108.