Marine accidents analysis based on data mining using K-medoids clustering and improved A priori algorithm

Baicheng Yang 1*, Zhilei Zhao1 and Jianbin Ma2
1 Navigation College, Dalian Maritime University, Dalian, Liaoning, 116026, China
2 Shenzhen Urban Public Safety and Technology Institute, Shenzhen, Guangdong, 510852, China
*Corresponding author’s e-mail: dmuybc@126.com

Abstract. In order to analyze the causes of maritime accidents and further ensure the safety of ship navigation, the risks of ships at sea and their impact factors are identified. The Zhejiang Sea area, China is used as the case study. Real maritime accident data of this area are collected and used as model inputs. The accident types are clustered center, and merges with the A priori algorithm to realize the classification mining of maritime accident data. 8 association rules of collision accidents were extracted under the conditions of 20% of the support threshold and 50% of the confidence threshold after removing negative association rules. The mining algorithm is based on K-medoids and A priori combination. K-medoids improves accuracy of results. In this paper, lift value is introduced to assist the evaluation of the association among these factors, which enhances the stability of results, compared with traditional methods. Finally, by analyzing the association rules one by one, the causes and characteristics of maritime accidents in Zhejiang waters are summarized.

1. Introduction
The number of coastal maritime accidents in China sea area has been declined in recent years, owing to the development of the intelligence of ship navigation system, the quality of crewmember, as well as the efforts put by maritime authorities of various countries. However, due to the complexity of navigation routes, the high density of ships, and the severe natural environment in certain sea areas, maritime navigation safety is still a significant issue. In order to find out the main causes of maritime accidents and mitigate ship navigation risks, researchers have conducted extensive studies on traffic accidents in many sea areas from different views and achieved substantial results. The current dominant research method is to assess safety of complicated sea area based on index system established by human factor-ship and cargo-environment-management. Fuzzy theory, grey relation theory, neural network algorithm are used to determine the risk level of different sea areas and identify main risks in target sea area. However, since there exist subjective factors in the establishment of the index system, and in some cases, less or no accident data can be used as the model inputs, there are certain doubts in the accuracy and navigation practice verification [1-5].

F Goerlandt [6] conducted a data visualization mining on natural environment and ship accidents in the north Baltic sea from 2007 to 2013. The causes of the maritime accidents in winter are identified. Moreover, the relationship between ship traffic accidents and sea ice, meteorology, and human operation are analyzed. Liu Zhengjiang [7] investigated the reports of nearly 100 ship collision accidents and extracted the causes of each accident. Association mining is conducted between human
errors and induced factors, as well as collision accidents. This work also determined the corresponding relationship between human errors and induced factors during ship collision accidents. Zhang Xiaohui [8-9] conducted an all-factor association mining experiment on water traffic accidents in various coastal areas of China. Ship collision risk is identified as the most prominent risk in the Yangtze river delta waters. Huang Changhai [10] established an association rule model and accident factor network. The author extracted 15 strong association rules by analyzing the support and confidence value in experiment results. A detailed analysis of the association factors of minor accidents was conducted. Goerlandt [6] and Liu Zhengjiang [7] did not analyze the connection between accident causes and its characteristics, such as accident locations and ship types. Zhang Xiaohui [8-9] and Huang Changhai [10] analyzed association rules base on the support value only. It is difficult to recognize noise in result. The stability is weak.

Author well-pleasing the above research results and in consideration of the differences of natural environment and traffic environment in various waters, further targeted analysis of maritime accidents in Zhejiang waters is carried out. Meanwhile, in order to avoid reduction of other risk factors and reduction of mining precision caused by the all-factor association mining of accident data, maritime accident in Zhejiang waters is the object of mining. We propose a combined mining method of K-Medoids cluster analysis and improved A priori association rules. Lift value is introduced into experiment. First, the accident was clustered and then the data after clustering was deeply mined. Eliminate negative association rules, with the value of appreciation as a rule of re-strengthening standards to extract 7 strong association rules of collision accidents. Proposed method further improved the precision and speed of mining experiment.

2. Maritime accidents data mining procedure

2.1. Data preparation
Data preparation mainly includes four processes: data collection, data preprocessing, data cleaning and data transformation [11, 12]. Original data of maritime accidents in Zhejiang waters were obtained by investigating and surveying. 526 samples of maritime accidents from 2008 to 2014 were selected and processed. Impact factors were transformed into code.

2.2. K-medoids clustering
As a first step, the accident database was clustered for the preparation of the deep mining of data. From a global and systematic point of view, this database is regarded as an integrity closed database. In this case, the K-medoids clustering algorithm is applicable. The algorithm is illustrated step by step in the rest of the section.

- Draw k objects at random from n accident data as the initial cluster center.
- According to the mean value of each cluster object (central object), the distance between each object and these central objects is calculated, corresponding objects are reclassified according to the minimum distance.
- The mean value (central object) of each (variable) cluster is recalculated.
- Calculate the standard measure function, if the function converges the algorithm terminates, if not back to process 2.

The error sum of squares criterion function is a common indicator used to evaluate the clustering performance in K-medoids clustering algorithm. Assume a mixed sample \( X = \{X_1, X_2, \ldots, X_n\} \), adopt a similarity measurement, \( X \) is clustered and grouped into \( k \) separate subsets \( X_1, X_2, \ldots, X_k \). Each subset is a type and contains samples \( n_1, n_2, \ldots, n_k \), respectively. In order to measure the quality of the cluster, the sum of error squares \( J_k \) clustering criterion function is used, which is defined as:

\[
J_k = \sum_{j=1}^{n_k} \sum_{i=1}^{n_k} \| x_{ij} - m_j \|^2
\]  

(1)
The accident database is taken as the cluster object. R language is used to conduct the K-medoids clustering, taking accident types and accident causes as the cluster center. PAM function is introduced to cluster the complete set of accidents.

2.3. A priori algorithm and improvements

A priori algorithm is a method to identify and analyze the underlying regularity of different sets in the database and usually divided into simple association, temporal association and causal association.

The clustered database is used as the basic database for mining. Define the accident database \( D \) and the item set \( \text{Item}=\{\text{Item}_1, \text{Item}_2, \cdots \text{Item}_n\} \) in the database. Thus, the association rule is the implication of the form \( A \Rightarrow B \), wherein \( A \subseteq \text{Item} \), \( B \subseteq \text{Item} \) and \( A \cap B = \emptyset \). The support of item set \( A \) represents the number of times that item set \( A \) appears in all item set \( I \), which is \( \text{sup}(A) = \frac{\text{TRTR} \supseteq A}{\text{|TR|}} \). So the confidence association rule \( A \Rightarrow B \) can be expressed as \( \text{conf}(A \Rightarrow B) = \frac{\text{sup}(A \cap B)}{\text{sup}(A)} \), lift can be expressed as:

\[
\text{Lift}(A \Rightarrow B) = \frac{\text{sup}(A \cap B)}{\text{sup}(A) \ast \text{sup}(B)}
\]  

(2)

The lift reflects the correlation between \( A \) and \( B \) in the association rules. When the lift is more than 1, the higher it is the higher the positive correlation. When the lift is less than 1, the higher it is the higher the negative correlation. When the lift is 1, there is no correlation. Since the A priori algorithm in setting up support and confidence thresholds often associated with the data provided by the user, the user continuously explores support and confidence thresholds to obtained reasonable and effective association rules. Analysis of the algorithm results also pays much attention to the pursuit of a higher support and confidence, and ignore the evaluated effectiveness of the rules by lift. However, in this paper, the association rule results are obtained through exploring the appropriate threshold of support and confidence firstly. A reasonable number of strong association rules are obtained, followed which effective strong association rules are obtained by removing redundant rules and negative association rules. Compared with the built-in software of other algorithms, R language has a great flexibility in handling the details of fixed algorithms. By programming in R language, frequent item sets are found in the database. At this stage, by setting the minimum support threshold and confidence threshold, the pruning process is carried out to generate the required strong association rules.

One of the major contributions of this work is the improvement of the A priori algorithm by realizing rapid extraction of the association rules when scanning huge database. To achieve this, a dynamic storage space base on \((k-1)\)-frequent item set is established. Each item in \((k-1)\)-frequent item set is linked when the \((k-1)\)-frequent item set is generated by the A priori algorithm. It can thus reduce scan number of times when frequent item set is generated. As a result, \(k\)-frequent item set could be extracted faster than the original, which improves the calculating speed.

3. Maritime accidents cluster and analysis

3.1. Accident cluster

Taking accidents in Zhejiang waters as an example, considering the accident characteristics, tonnage of ship, and accidents type are taken as the clustering centers to realize the dynamic clustering of the database by using R language. According to the clustering results, by accidents type, the database is clustered into two categories, collision and non-collision accidents. After the completion of the accident clustering, the clustering and association are visualized through the association among various types. The accident type-oriented network graph and association rule distribution scatter diagram are generated and shown in figure 1. Considering the sample capacity of the accident database, and the precision of the experiment, the accidents are grouped into two categories. It is noted that, the support value of subset in database is improved by clustering, which also improves the mining accuracy indirectly.
Figure 1. The maritime accidents clustering

3.2. Accident information network analysis

Taking collision accidents as an example, 179 effective data of collision accidents are obtained through K-medoids clustering of maritime accidents database in Zhejiang waters. The collision-oriented network diagram is generated based on the accident types and can be used for rough analysis even though the detailed association rules cannot be obtained. Considering the limitation of the number of nodes, links, and size of graph, the product of the support threshold and the total data sample is taken as the link threshold. By doing so, the frequent candidate set is reserved, moreover, the level of strength of each association rule is accurately and intuitively expressed. Through analyzing the collision accidents in the database, adopting the support threshold of 20%, the collision-oriented network diagram is generated and shown in figure 2. Several observations could be made:

- Negligence of looking out, failure to use safe speed, restricted visibility and improper collision avoidance behavior present high frequency of occurrence in factors caused collision accidents.
- From a view of accidents occurred location, the collision accidents mainly occurred in the coastal waters. As for the ship types, fishing ships, sand carriers, and general dry cargo vessels are mainly involved in collision accidents.
- In terms of ship tonnage, ships less than 3000GT (Code is tonnage-Ⅰ) are more likely to occur collision accidents.
- Regarding the analysis of time series, the time period from 8 pm to 12 pm is the most frequent time period of collision accidents in Zhejiang waters. According to the analysis of the seasonal series, spring is the main season for collision accidents, accounting for more than 40% of the total accidents, followed by summer.
- With regards to the damages caused by collision, most of the economic losses caused by the collision are less than 1 million RMB (Code is economic-Ⅰ/Ⅱ).
4. Maritime accidents mining and analysis

4.1. Accidents mining

In this paper, 474 collision accident association rules and 304 non-collision accident association rules are generated during the classification of the data mining ship traffic accidents association rules in Zhejiang waters. Taking collision accidents as an example, the scatter diagram of all the association rules is shown in figure 3.

Figure 3. The scatter plot collision accident association rules distribution

Based on the adjustment of the support threshold and confidence threshold, eight collision accident association rules with lift value more than 1.4 were selected. The conditions of 20% support threshold and 50% confidence threshold of collision accidents are determined according to the ranking of lifting
appreciation. Collision accident association rules are selected as an example for analysis, listed in Table 1.

## Table 1. Maritime accident association rules of Zhejiang waters

| NO | Antecedent | Succedent | Sup(%) | Conf(%) | Lift |
|----|------------|-----------|--------|---------|------|
| 1  | type-fishing ship; tonnage≤3000GT; class 1 cause- human factor | | 23.32 | 91.72 | 1.58 |
| 2  | type-sand carrier; tonnage≤3000GT; class 1 cause- human factor | | 21.57 | 98.01 | 1.58 |
| 3  | type-sand carrier; improper collision avoidance behavior | | 21.66 | 74.42 | 1.54 |
| 4  | type-fishing ship; improper collision avoidance behavior; tonnage≤3000GT; class 1 cause-human factor | | 23.12 | 72.39 | 1.5 |
| 5  | type-fishing ship; accident area-coastal | | 22.13 | 78.06 | 1.49 |
| 6  | type-sand carrier; improper collision avoidance behavior; accident level-serious | | 24.32 | 72.01 | 1.46 |
| 7  | season-spring | | 25 | 71.22 | 1.41 |

### 4.2. Mining results, discussion and implications

On the basis of maritime accidents in Zhejiang waters, the mapping relationship between single accident characteristics is discovered by data mining and clustering. The identification of potential rules among multiple factors is also carried out. The lift value is introduced to be an important evaluation attribute, and to correspond to the support and confidence value. Three attributes validate each other. Noise of the results is reduced. The analysis of association rules getting from this experiment is as follows:

- 91.72% accidents are related to human factors in fishing ships collision accidents.
- Also regarding fishing ships less than 3,000 GT, 78.06% accidents occurs in coastal waters and the cause of the accidents is related with human factors.
- As for sand carriers less than 3,000GT, 98.01% collision accidents are related to human factors.
- The improper collision avoidance behavior is the main cause leading to sand carrier collision, accounting for 72.01% of all the accidents. At the same time, the damages of such accidents are serious in term of economic loss.
- When ship collision occurs in spring, 71.22 % is associated with restricted visibility.

The potential regulars of the above traffic accidents are obtained by cluster mining. The correlation among the attributes, time series, accident causes, and ship types are identified. Based on this method, the maritime safety authorities and shipping companies could formulate targeted safety measures. Protect priority could be identified, and efficient resource allocation is thus achieved.
5. Conclusion
The presented work combines data mining with association rules algorithm and clustering algorithm. It introduces lift value and improved A priori algorithm. Accident data in Zhejiang waters is used as the case study. Control of algorithm details is more rigorous through R language. Without reducing the threshold of support and confidence obtained in previous studies, the association rules can be effectively positive strengthened by the lift value. The theoretical results support the analysis of practical experience and work well. Its relevance and application value is better than previous studies. The stability of results is better.

In future studies, maritime accident data covering longer research period could be adopted. The accidents could also be studied according to the seasons. On the other hand, more cluster centers, apart from the accident type, could be considered, such as the damages of accidents and recovery periods. The proposed framework shows its applicability in other waters and against other hazards as well.

References
[1] Li, H.X. (2014) Study on the data-mining based mechanism for the occurrence of ship collision accidents caused by human factors. information technology for manufacturing systems. Applied Mechanics and Materials, Vol 686: 321–326.
[2] Dimitrios, X., Dionisios, S. (2011) An application of data mining tools for the study of shipping safety in restricted waters. Safety Science, Vol 49: 192–197.
[3] Wang, F.W., Wu, Z.L., Qi, Z. (2013) A study of safety assessment of ship to sail based on rough set theory. Navigation of China, Vol 02: 61–66.
[4] LI, B. (2010) Pre-evaluation on navigation safety in tianjin port compound seaway[D]. Dalian Maritime University.
[5] Yang, G.B. (2006) The research of computing the vessels’ collision risk index based on neural network with genetic algorithm [D]. Hebei University of Technology.
[6] Goerlandt, F., Goite, H., Osiris, A. (2017) An analysis of wintertime navigational accidents in the Northern Baltic Sea Safety. Science, Vol 92: 66–84.
[7] Liu, Z.J., Wu, Z.L. (2004) Data mining to human factors based on ship collision accident survey reports. Navigation of China, Vol 02: 3–8.
[8] Zhang, X.H., Liu, Z.J., Wu, Z.L. (2011) Pattern discovery based on MAMIS. Journal of Dalian Maritime University, Vol 02: 39–42.
[9] Zhang, X.H. (2011) Applications of cloud theory and data mining in marine safety analysis [D]. Dalian Maritime University.
[10] Huang, C.H., Gao, D.Y., Hu, S.P. (2014) Association rule analysis of vessel traffic accidents based on A priori algorithm. Journal of Shanghai Maritime University, Vol 03: 18–22.
[11] Ma, J.B., Chen, H.Y. (2016) The association analysis of vessel traffic accident based on data mining. Ship&Ocean Engineering, Vol A01: 66–67.
[12] Chen, Y., Li, T.Y. (2012) Data Mining and Clustering Analysis. Dalian Maritime University Press Co. Ltd., Dalian.