A Target Behavior Law Mining Method Based on Similar Duplicate Record Detection

Xueteng Li, Hualin Liu, Youpeng Huang, Haipeng Wang, Xiangqi Gu, Hao Liu, Xinlong Pan and Shuyi Jia

ABSTRACT

In the complex marine environment, with the continuous enrichment and development of target detection methods, massive target track data is stored and accumulated in the database, which contains a large amount of information and knowledge. It is possible to analyze and predict the behavior and intention of the target by means of target behavior law mining. Most of the existing target behavior law mining methods are based on the idea of clustering. When working in the complex environment of the sea battlefield, the working steps are complicated and the effect is poor. Therefore, in combination with situation analysis requirements and based on the idea of similar duplicate record detection, this paper proposes a mining method of maritime target behavior law by defining multi-dimensional record matching similarity (MDRS). The experimental analysis is carried out on the simulated military scene. The results show that the proposed method can accurately and effectively mine the behavior of the target.

1. INTRODUCTION

The target behavior law refers to the similar or identical behavioral characteristics of the target when performing specific tasks in the past, such as the
same or similar route, speed, course and other behavioral characteristics of the
target. At present, most trajectory data mining techniques are implemented by
clustering algorithm. However, in view of the many interference factors and large
data volume in the sea battlefield environment, the clustering algorithm has the
following disadvantages: First, the clustering algorithm mostly converges the
trajectories with similar spatial positions into one class, and rarely has trajectory
clustering methods that consider the multi-dimensional features of the position,
velocity, heading, attributes and types of the target required for battlefield situation
analysis; Second, the working process of the clustering algorithm is complicated,
and the large amount of data in the sea battlefield will slow down the running
speed; Third, the clustering algorithm is sensitive to the setting of parameters and
the target track in the sea battlefield is complex and variable. It is difficult to unify
the algorithm parameters in face of different types of targets. Finally, when using
the clustering algorithm, some noise points that do not belong to any class are often
separated. In the sea battlefield environment, these noise points are likely to be
special groups with single elements. If these special groups are treated as noise
points, it may have a big impact on the results of the battlefield situation analysis.

To solve this problem, this paper refers to the repeated data cleaning
technology [1] in data pre-processing technology, and regards each target track
data as a record. In this way, when processing the multi-dimensional track data, the
two tracks with similar behaviors (similar positions, speed, course and the same
attributes, type) can be regarded as similar records, which transforms the complex
problem of target track behavior law mining[2-3] into a relatively simple problem
of similar duplicate record detection. In addition, in the traditional similar duplicate
record detection [4-5], multi-source data is usually string data containing
identifiers. Whether two strings are equal cannot be obtained only by arithmetic,
and it is often necessary to define a set of equivalence rules, which makes the
cleaning of repeated records very complicated. However, in the sea battlefield
environment, the multi-source data is the message data formed after special coding,
which means the corresponding attribute information in the track is expressed in
numerical form (such as the target attribute: using "0", "1", "2" to represent our
side, the enemy and the middle cube). Therefore, applying similar duplicate
recording detection technology to the sea battlefield data processing process can
greatly reduce the complexity of matching similarity between computational
records, and is conducive to the mining of target track behavior law.

2. MULTI-DIMENSIONAL TRACK DATA OF MARITIME TARGETS

In the process of maritime situation analysis, the target track data is usually a
multi-dimensional sequence composed of multi-dimensional track data points [6],
including time, position, speed, course, attribute, type and other multidimensional
information [7].
Set the track sample data set in the target area as

\[ TR = \{TR_1, TR_2, \ldots, TR_n\} \]  \hspace{1cm} (1)

Among them, \( TR \) is the set of all track samples, \( i \in [1, n] \) is the sample number, \( n \) is the total number of sample data set. Each target track \( TR_i \) contains \( m \) chronological multidimensional data points, i.e.,

\[ TR_i = \{TR_i(1), TR_i(2), \ldots, TR_i(j), \ldots, TR_i(m)\} \]  \hspace{1cm} (2)

Among them, \( TR_i(j) \) is the \( j \)th multi-dimensional track data point in the \( i \)th track sample, \( j \in [1, m] \) is the track data point number, \( m \) is the total number of sample data points. Each track data point \( TR_i(j) \) contains \( q \) multi-dimensional features, its expression is

\[ TR_i(j) = \{ \text{time}, \text{position}, \text{velocity}, \text{acceleration}, \text{course}, \text{attribute}, \text{type} \ldots \} \]  \hspace{1cm} (3)

Among them, \( TR_i(j) \) has \( q \) elements, represents that the \( j \)th multi-dimensional track data point in the \( i \)th track sample contains \( q \) multi-dimensional features such as time, position, speed, acceleration, course, attribute and type.

3. TARGET BEHAVIOR LAW MINING METHOD

3.1 Overview of Similar Duplicate Record Detection

When building a database, a large amount of multi-source data needs to be imported and fused. Ideally, each entity in the real world will have a unique record in the database.

![Figure 1. The flow chart of similar duplicate record detection.](image-url)
However, in practice, if the records of the same entity from different data sources are different, and the system cannot effectively identify and determine that they are the same entity, the merged database will retain both records. If two similar but incompletely repeated records in the database represent the same entity, the two records are said to be similar records. If there are two records in the database with identical attribute values, they are called duplicate records. Generally, similar records and duplicate records are collectively referred to as similar duplicate records, that is, multiple records in the database corresponding to the same entity.

Similar duplicate record detection is the process of finding records representing the same entity in the real world from a multi-source data set. To judge whether the two records are similarly repeated, first compare the corresponding attributes of the records and calculate their field similarity. Then, according to the weight of each attribute, performing the weighted average to obtain the record similarity. If the similarity of two records exceeds a certain threshold, the two records are considered to be matched and similarly repeated. In summary, the basic process is shown in Fig.1.

Based on the idea of similar duplicate record detection, the article considers each target track data TR in the data set TR as a record, then m data points constituting the target track TR are the m sub-records of the record. The q features such as time, position, speed, acceleration, course, attribute and type contained in each data point TR(j) are q fields of the sub-record.

3.2 Method Flow

Based on the above thoughts of similar duplicate record detection, this paper proposes a method of mining the maritime target behavior law, which processes the historical track data in the historical database, screens out the target with the law of behavior and the target with irregular behavior in the target area, and classified storage. The specific process is as follows:

Step 1: Compare the corresponding attributes of the record. There are many attributes in target track data, so it is necessary to select attributes that can represent the record characteristics while the record matching, which requires the operator to have a deep understanding of the data meaning, and select appropriate attributes based on battlefield requirements and expert experience.

Step 2: Calculate the field similarity (FS). Field matching [8-9] is the basis of record matching and the core issue of similar duplicate record detection. FS is a numerical value calculated based on the field information to indicate the degree of similarity between two fields, and 0 ≤ FS ≤ 1. The larger FS is, the more similar the two fields are. If FS = 1, the two fields are duplicate fields. For the numerical field in the maritime operational environment, set the field values of TR₁ and TR₂ at the same time point and for the same attribute x as x₁ and x₂, then
\[
FS(x_1, x_2) = \left(1 - \frac{|x_1 - x_2|}{\max(x_1 - x_2)}\right) \times 100\%
\]  

(4)

Step 3: Assign attribute weight and Imperium value. The weight of attribute \[10\] represents the importance of an attribute in determining the similarity between two records. The more important the attribute is, the bigger weight is assigned. The sum of all attribute weights is 1.

Aiming at the maritime combat environment, this paper proposes the concept of \(\text{ImValue}\) to optimize the comprehensive weight. Let \(TR_1\) and \(TR_2\) have similar position, speed and course characteristics at the same time, but they have different attributes, so \(TR_1\) and \(TR_2\) cannot be similar duplicate records. It can be concluded that only when certain attribute values of two records completely match can the records be compared. Set the weight of the attribute as \(\text{ImValue} = 1\), which is called Imperium value. the field similarity of the attribute corresponding to the Imperium value can only be 0 or 1. The method can effectively improve the accuracy of similar duplicate record detection, reduce the number of field matching and shorten the running time.

When calculating the record similarity, the attribute value corresponding to the Imperium value is multiplied by the attribute value corresponding to the traditional weight.

Step 4: Calculate the multi-dimensional record similarity (\(MDRS\)). If the \(MDRS\) of two records exceeds a certain threshold, the two records are considered to be matched and point to the same entity, in another word, they are similar duplicate records.

Let each record \(TR_i\) have \(p + q\) attribute fields for comparison \(Field(1), Field(2), \ldots, Field(p + q)\). There are \(q\) attributes corresponding to Imperium value, so the attribute weight of each field is \(W(1), W(2), \ldots, W(p), \ldots, \text{ImValue}(1), \text{ImValue}(2), \ldots, \text{ImValue}(q)\), which satisfies

\[
\begin{align*}
\sum_{i=1}^{p} W(i) &= 1 \\
\text{ImValue}(j) &= 1, \quad j = 1, 2, \ldots, q
\end{align*}
\]  

(5)

This paper uses an efficient, application-independent \textit{Pair-wise} comparison algorithm \([11-13]\), which is described as follows:

First, Given two records \(TR_1\) and \(TR_2\), Let \(FS(l), FS(2), \ldots, FS(i), \ldots, FS(p + q), i \in [1, p + q]\) be the field similarity of each field of the sub-record \(TR_i(r)\) and \(TR_2(r)\), \(r \in [1, m]\). If any field similarity of \(TR_1(r)\) and \(TR_2(r)\) less than the
threshold $\delta_i$, that is, $FS(i) < \delta_i$, then $TR_i(r)$ and $TR_2(r)$ are not similar duplicate records.

Second, according to the field similarity $FS(1), FS(2), \ldots, FS(i), \ldots, FS(p+q)$ and the corresponding attribute weight, the weighted average is carried out to calculate the records similarity of sub-records $TR_i(r)$ and $TR_2(r)$, i.e.,

$$RS(TR_i(r), TR_2(r)) = \frac{\sum_{i=1}^{p} W(i) \cdot FS(i) \prod_{j=1}^{q} ImValue(j) \cdot FS(j)}{\sum_{i=1}^{p} W(i)}$$ \hspace{1cm} (6)

By averaging the record similarity of all $m$ sub-records, the $MDRS$ of records $TR_i$ and $TR_2$ can be obtained as follows,

$$MDRS(TR_i, TR_2) = \frac{\sum_{r=1}^{m} RS(TR_i(r), TR_2(r))}{m}$$ \hspace{1cm} (7)

If $MDRS(TR_i, TR_2)$ is greater than the threshold $\delta_2$, $R_i$ and $R_2$ are considered to be similar duplicate records.

Among them, $\delta_i$ is defined as the similarity threshold of two fields, and $\delta_2$ is defined as the similarity threshold of two records.

Step 5: Detect similar duplicate records. In view of the demand of maritime combat situations, from the perspective of precision and recall rate, the best way to detect similar duplicate records in this environment is to match and compare each record $TR_i$ with all other records $TR_j (j \neq i)$ in the multi-source data set one by one[14].

Step 6: Merge or delete. After the similar duplicate record detection is completed, similar records are merged, only one correct record is retained, and other records are deleted. Generally, the combination of manual and automatic is adopted. Most of the similar repeat records are automatically combined by preset rules (such as random rules, latest rules and comprehensive rules), while a small part is manually completed. From this, we can select the target set $C_{TR}$ with behavior law and the target set $Anom_{TR}$ with behavior irregularity.

4. EXPERIMENTAL VERIFICATION ANALYSIS

In order to verify the performance of the target behavior law mining method proposed in this paper, experiments are carried out on the historical simulation data set of a military scene and the results are analyzed. The simulated
multi-dimensional track data is used to simulate the activity of the battlefield target, and the target track data is processed.

Figure 2. Plot of all trajectories in historical simulation dataset.

4.1 Data Set

This simulation data set assumes that the battlefield environment simulation is performed in an area of 450×350 square nautical miles in a wide sea from 12 to 18 in the past N days. The data set includes 109×N multi-dimensional target trajectories with behavior law and anomaly multi-dimensional target trajectories with irregular behavior. Among the 109 targets with behavior law, there are 69 fixed islands in 8 fixed island groups, 6 floating objects at sea, 11 civilian vessels, 3 our military formations contain 12 target vessels, and 3 enemy military formations contain 11 target vessels. Assume that the battlefield situation information is updated every 10 minutes, and the sailing height of the vessel in the sea area is relatively fixed. Therefore, each target track data is composed of 36 data points without considering the altitude information of the target. Based on the traditional two-dimensional position information and combined with the reality of the modern sea battlefield, each target track data is artificially added with multi-dimensional information such as date, initial velocity/acceleration, attribute and type to calculate the speed, position and course characteristics of the target. Finally, all tracks are randomly sorted to construct a (109×N+anom) ×11×36 multi-dimensional target track history simulation data set $P_{TR}$, as shown in Fig.2.
4.2 Experiment analysis

The historical simulation data set shown in Fig.2(b) is selected as the experimental scenario. Enter the appropriate field similarity threshold $\delta_1$ and record similarity threshold $\delta_2$ (in this experiment, take $\delta_1 = 0.9$, $\delta_2 = 0.97$). Based on the idea of similar duplicate record detection, the behavior law of historical target tracks is mined, and the output is shown in Fig.3.

It should be noted that when calculating $MDRS$ in this simulation environment, the Imperium value are set for the date, attribute and type fields. That is, if the two records have the same date, the field comparison cannot be performed; if the attribute or type of the two records are different from each other, the field comparison cannot be performed.

Among them, Fig.3(a) is the plot of all trajectories of behavior law, which including fixed targets and regular sailing targets; Fig.3(a) is the plot of all trajectories of irregular behavior, and such targets may be abnormal track targets due to their low frequency, which should be focused on and targeted.

The experiment evaluates the quality of target behavior law mining by calculating the recall rate ($R$) and precision rate ($P$) \[15\].

The recall rate is the percentage of the number of targets that are correctly mined and the actual number of targets, i.e.,

$$R = \frac{\text{the number of correctly mined targets}}{\text{the number of targets}} \times 100\%$$ \hspace{1cm} (8)

The precision rate is the percentage of the number of targets that are correctly mined and the number of targets that are mined, i.e.,
According to formulas (8) and (9), the plot of Recall and Precision dependent on the value of dataset size is shown in Fig.4. Among them, the horizontal axis of Fig. 4 represents the change of dataset size, and the numerical value is a multiple of the data amount in the simulation data set shown in Fig.4.

Fig.4(a) shows the curve of the Recall as a function of the data set size, and Fig.4 (b) shows the curve of the Precision as a function of the data set size. It can be seen from the figure that in this military scenario, with the increasing scale of the data set, the target behavior law mining method proposed in this paper can still accurately and efficiently realize the target behavior law mining in the maritime combat environment.

5. CONCLUSIONS

Aim at the demand of maritime combat situation generation, this paper combines the data cleaning technology with the target behavior law mining, designs a maritime target behavior law mining method based on the defined MDRS . The experimental verification was carried out on the simulated military scene, and the results were evaluated and analyzed by calculating the recall rate $R$ and the accuracy rate $P$. The results show that this method can effectively mine the behavior of historical target track and improve the efficiency and accuracy of sea situation generation. However, since the data set to be processed is large, and the complexity of the record similarity matching is $N(N-1)/2$ (N is the number of records in the data set), which is a relatively complicated calculation process, the method takes a long time.
ACKNOWLEDGMENTS

The authors would like to thank Prof. Xueyuan Lin, Dr. Qiang Guo and Ms. Meng Wang for their selfless help, support and encouragement in the paper writing process.

REFERENCES

1. Galbardas H, et al. 1999. “An Extensible Framework for Data Cleaning. Institute national de Recherche an Information et an automatous,” Technical Report 1999.
2. Xinlong Pan, You He, Haipeng Wang, Wei Xiong, Xuan Peng. 2016. “Mining regular behaviors based on multidimensional trajectories,” Expert Systems with Applications, 2016, 66(12): 106-113.
3. Xinlong Pan, You He, et al. 2017. “Online classification of frequent behaviors based on multidimensional trajectories,” IET Radar, Sonar & Navigation, 2017, 11(7): 1147-1154.
4. Jingyu Han, Lizhen Xu, Yisheng Dong. 2008. “An Overview of Data Quality Research,” Computer Science, 2008(02):1-5+12.
5. Cui Yong, Wu Jianping, Xu Ke, Xu Mingwei. 2002. “Research on Internetwork QoS Routing Algorithms: a Survey,” Journal of Software, 2002(11):2076-2082.
6. Pan Xinlong, Wang Haipeng, He You, Xiong Wei, Zhou Wei. 2017. “Anomalous behavior detection method based on multidimensional trajectory characteristics,” Acta Aeronautical et Astronautical Sinnica, 2017, 38(04): 254-263.
7. Yang A, Tan X, Baek J, et al. 2017. “A new ADS-B authentication framework based on efficient hierarchical identity-based signature with batch verification,” IEEE Transactions on Services Computing, 2015, 10(2): 165-175.
8. Bilenko M, Mooney R, Cohen W, et al. 2003. “Adaptive name matching in information integration,” IEEE Intelligent Systems, 2003, 18(5):16-23.
9. Monge A E, Elkan C P. 1996. “The Field Matching Problem: Algorithms and Applications,” In: Proceedings of the 2nd Conference on knowledge Discovery and Data Mining, Portland, Oregon. 1996:267-270.
10. SHE Chunhong. 2002. “Data Cleaning Approach,” Computer Applications, 2002, 12, 22 (12): 128-130.
11. Wagner R A, Fischer M J. 1974. “The string-to-string correction problem,” ACM, 1974, 21(1): 168-173.
12. Lowrance R, Robert AWagner.1975. “An extension of the string-to-string correction problem,” ACM, 1975, 22(2): 177-183.
13. Sellers P H. 1974. “On the theory and computation of evolutionarydistances. SIAM,” Applied Mathematics, 1974, 26(4): 787-793.
14. Hernandez M A, Stolfo S J. 1998. “Real-World Data is Dirty: Data Cleansing and the Merge/Purge Problem,” Data Mining and Knowledge Discovery,1998,2(1):9-37.
15. LIU Hua. 2007. The Research of Data Cleaning in Web Information Integration. Wuhan: Wuhan University of Technology, 2007.