Trajectory Clustering Of Segmented Field Operations Logistics Process

B.S.Zhang¹,a,* , Z.Q.Xiang²,b,* , H.Zhang³,c,* , R.Liu⁴,d
¹ dept. name of organization: School of Naval Architecture, Ocean and Energy Power Engineering º Wuhan, China * Weihai Research Institute of Wuhan University of Technology
² dept. name of organization: School of Naval Architecture, Ocean and Energy Power Engineering b Wuhan, China * Weihai Research Institute of Wuhan University of Technology
³ dept. name of organization: Jiangsu Automation Research Institute c Lianyungang, China
⁴ dept. name of organization: School of Naval Architecture, Ocean and Energy Power Engineering d Wuhan, China
* zh397312701@163.com

Abstract—It has become a new research direction to use process mining technology to mine and analyze all kinds of data generated during shipbuilding. The classification of hull segments into different clusters according to some properties in advance can make the excavation results more specific. In this paper, the process examples in ship construction are divided into different clusters based on the cohesive hierarchical clustering algorithm. Firstly, this paper introduces the advantages and disadvantages of several main clustering algorithms, selects the clustering algorithm most suitable for segmented outftield logistics, establishes the mathematical model of segmented outftield logistics process trajectory clustering, and defines the appropriate feature vectors. Euclidean distance, Hamming distance, Jekard distance and cosine distance were used to calculate the similarity distance between process instances, and the concepts of chain, single chain and group average were added to select the appropriate similarity distance between clusters. An evaluation method of clustering results based on contour coefficient is introduced which can effectively evaluate the clustering results of the logistics process trajectory during segmented remote operations. Finally, the feasibility and effectiveness of the proposed algorithm are verified by experiments.

1. Introduction
Clustering (Clustering) is in accordance with certain attributes of data objects, the initial set of data objects can be divided into different clusters (Cluster), each Cluster is a subset of the initial set of data objects, the data of data objects in the group of the object with the same as similar as possible, and other data in a Cluster objects as far as possible.

There are many clustering algorithms in the field of data mining, mainly including hierarchical clustering, partitioning clustering, density-based clustering, grid-based clustering and model-based clustering. In recent years, methods like quantum clustering, spectral clustering have gradually become popular. Seungkyung Lee used process mining technology to analyze the manufacturing process of the
segment after assembly. He used agglomerative hierarchical clustering algorithm to cluster the segment transportation event log. He introduced the correlation coefficient about the same appearance to select the optimal similarity calculation method. Dongha Lee conducted clustering from the perspectives of job task and job shop respectively. He integrated the two groups of clustering results into a new group of clusters by using a mixed matrix, and extracted the process model according to the clustering results.

In this paper, for the event log in the logistics process of segmented outfield operations, each event has a timestamp attribute, so there will be a sequential relationship between different events. Some clustering algorithms in the field of data mining are applied directly, so it is difficult to take into account the timestamp attribute and sequence relationship in the event log data. Therefore, the traditional data clustering is converted into trajectory clustering, so as to meet the clustering requirements of event log in the segmented outfield logistics process.

2. Trajectory clustering
Due to the poor performance of traditional process mining algorithm in dealing with unstructured event logs with a large number of process instances and a great difference between different process instances, Trace Clustering is generated. Each trace in the trajectory clustering is a process instance in the event log, which is an actual process, and the clustering of the trajectory is the clustering of all the process instances in the event log. As shown in Figure 1, the basic idea of trajectory clustering is to divide all process instances in the event log into different clusters according to some properties of process instances in the event log, so as to ensure higher similarity between process instances in the same cluster and lower similarity between process instances in different clusters.

![Fig 1 Diagram of trajectory clustering](image)

In the event log used by trajectory clustering, each event has a timestamp attribute, and there is a sequential relationship between different events. However, most of the data used by clustering algorithms in the data mining field are point-like data, and there is no such time-stamp attribute and sequential relationship between data points. Therefore, a multi-dimensional feature vector is introduced to transform the basic data clustering algorithm into a trajectory clustering algorithm that can cluster the event log, so as to ensure that each process instance in the event log has a corresponding feature vector. The eigenvector is used to calculate the similarity distance between process instances. Active attributes, timestamp attributes, resource attributes, and sequential relationships can all be used as angles in multidimensional eigenvectors.

Due to the small batch production of the hull block, it is difficult to pre-determine reasonable initial clustering centers for various types, and there is great uncertainty in randomly initializing clustering centers. The process examples of different segmented field operations logistics are very different. The density is very uneven, and this kind of data set is generally large and high-dimensional. Therefore, the most widely used hierarchical clustering algorithm (AGNES) is selected as the basic clustering algorithm for the event log trajectory clustering of piecewise outfield job logistics process. It can avoid the defects of other algorithms when using hierarchical clustering algorithm to carry out track clustering of event logs in segmented outfield job logistics process. Although hierarchical clustering also has some defects, which have certain influence on clustering efficiency and results, it is reasonable to take hierarchical clustering algorithm as the basic clustering algorithm.
3. Mathematical modeling

3.1. The feature vector of the event log of the segmented outfield operation logistics process

Multi-dimensional feature vectors are the core of trajectory clustering algorithm. In this section, feature vectors are defined and integrated respectively from three perspectives of job task type, job site and job order.

First, suppose the number of process instances in the event log of segmented field operations logistics process is \( N \). Existing process instance \( i, j \in [1:N] \), but \( i \neq j \).

When defining eigenvectors from the perspective of job task type, if the number of job tasks is \( m \), each Instance\( i \) will generate an eigenvector with dimension \( m \) here. If the eigenvector is \( \alpha_i \), then the eigenvector for job task type is defined as:

\[
\alpha_i = (a_{1,i}, a_{2,i}, \ldots, a_{m,i})
\]

(1)

\( a_{(1,\ldots,m)} \) is an element in the eigenvector, and each element value represents the number of times the corresponding job task of the element is executed. The element values in \( a_{(1,\ldots,m)} \) are usually 0, 1, or 2.

When defining the eigenvector from the perspective of the Instance\( i \) operation site, let the number of operation sites be \( n \), then each Instance\( i \) will generate an eigenvector with dimension \( n \), and let the eigenvector be \( \beta_i \). In terms of Instance\( i \) operation site, the eigenvector is defined as:

\[
\beta_i = (b_1, b_2, \ldots, b_n)
\]

(2)

Where, \( c_{(1,\ldots,m')} \) is an element in the eigenvector, and each element value represents the operation times of the corresponding operating site of the element. The value in \( b_{(1,\ldots,n)} \) is generally 0 or 1. If the task is reworked and there are multiple transfers between the two sites, the element value may reach 2, but it is generally no higher.

When defining the eigenvectors from the perspective of job order, since the number of job types is set as \( m \), and each job is combined in pairs, \( m^2 \) order relations will be generated. Every instance\( i \) will generate an eigenvector with dimension \( m^2 \) here. If the eigenvector is \( \gamma_i \), the eigenvector in job order is defined as:

\[
\gamma_i = (c_{1,i}, c_{2,i}, \ldots, c_{m^2,i})
\]

(3)

Where, \( c_{(1,\ldots,m')} \) is an element in the eigenvector, and each element value represents whether there is a job order relationship corresponding to this element in instance. \( c_{(1,\ldots,m')} \) of the element value and \( a_{(1,\ldots,n)} \), \( b_{(1,\ldots,n)} \) in the element values are not the same. Each element value can only be taken as 0 or 1. When the element value is 0, the corresponding job order relationship of the element exists in the representation; otherwise, it does not exist.

Finally, feature vectors of job type, job site and job order are integrated together to form a multi-dimensional feature vector of event log of segmented outfield job logistics process with a dimension of \( m+n+m^2 \). Assuming this feature vector is \( \epsilon_i \), the overall feature vector of Instance\( i \) is defined as:

\[
\epsilon_i = (\alpha_i, \beta_i, \gamma_i) = (a_{1,i}, a_{2,i}, \ldots, a_{m,i}, b_1, b_2, \ldots, b_n, c_{1,i}, c_{2,i}, \ldots, c_{m^2,i})
\]

(4)

The element of \( \epsilon_i \) is reset as \( l \), and the characteristic vector \( \epsilon_i \) is finally updated as:

\[
\epsilon_i = (l_1, l_2, \ldots, l_{m+n+m^2})
\]

(5)

3.2. Calculation of similarity distance between process instances

Many similarity distance calculation methods are used in hierarchical clustering, such as Euclidean distance, Hamming distance, Jeckard distance, cosine distance, etc. Combined with the application
environment of track clustering in segmented outfield operation logistics process, the calculation formula of these similarity distance calculation methods is redefined.

3.2.1. Euclidean distance
Euclidean distance is derived from the straight-line distance formula between two points in space. Each eigenvector in the logbook of segmented outfield operation logistics process can be approximately regarded as a point in space, the Instance and Instance Euclidean distance is the linear distance of their corresponding eigenvectors in space. The calculation formula of the distance is defined as [1]:

$$ ED(i, j) = \sqrt{\sum_{q=1}^{m+n+m} \left( e_{iq} - e_{jq} \right)^2 } $$  \hspace{1cm} (6)

3.2.2. Hamming distance
Hamming distance refers to the number of vectors with different corresponding positions in two vectors of the same length, that is, the minimum number of changes needed to convert one vector into another. Hamming distance calculation formula is defined as follows:

$$ HD(i, j) = \sum_{q=1}^{m+n+m} \delta(e_{iq}, e_{jq}) $$ \hspace{1cm} (7)

$$ \delta(e_{iq}, e_{jq}) = \begin{cases} 0, & \text{if } (e_{iq} > 0 \wedge e_{jq} > 0) \vee (e_{iq} = e_{jq} = 0) \\ 1, & \text{otherwise} \end{cases} $$ \hspace{1cm} (8)

3.2.3. Gerard distance
The object to which The Gerkard distance applies is not a vector, but a set, which represents the proportion of all the elements in two sets of different elements. When applying The Jekard distance, the eigenvectors need to be expressed in the form of a set. However, the elements in the eigenvector whose element value is 0 do not exist in the set, and all the elements in the set are greater than 0 in the corresponding eigenvector. Therefore, Gerard's distance calculation formula is defined as [2]:

$$ JD(i, j) = \frac{\left| e_{iq} \cup e_{jq} \right| - \left| e_{iq} \cap e_{jq} \right|}{\left| e_{iq} \cup e_{jq} \right|} $$

$$ = 1 - \frac{m+n+m}{\sum_{q=1}^{m+n+m} e_{iq} e_{jq}} $$ \hspace{1cm} (9)

3.2.4. Cosine distance
The cosine of the Angle between two vectors measures the similarity between two data objects. The smaller the Angle between the two vectors is, the closer it is to 0 degree, and the closer the cosine is to 1, the more similar the two data objects are. The larger the included Angle is, the closer it is to 180 degrees, and the closer the cosine is to -1, the less similar the two data objects are. The cosine distance calculation formula is defined as [3]:

$$ CD(i, j) = \frac{\sum_{q=1}^{m+n+m} e_{iq} e_{jq}}{\sqrt{\sum_{q=1}^{m+n+m} e_{iq}^2} \sqrt{\sum_{q=1}^{m+n+m} e_{jq}^2}} $$ \hspace{1cm} (10)
4. Intercluster similarity
The ultimate goal of the clustering algorithm is to merge the clusters with the minimum similarity distance, rather than merge the process instances with the minimum similarity distance. There are multiple process instances in a cluster, and many similarity distances are generated between process instances. Therefore, methods such as SL, Single Link, CL, Complete Link and GA, Group Average were adopted to take one of the similarity distances as the similarity distance between clusters.

First define a set of cluster \( C = \{c_1, c_2, \ldots, c_k\} \), given two clusters \( C_i \) and \( C_z \), \( C_i, C_z \in C \). The similarity distance between is expressed as \( distance(C_i, C_z) = d(C_i, C_z) \).

4.1. Single Link
The minimum similarity distance between process instances in \( C_i \) and \( C_z \) clusters is selected as the similarity distance between clusters, which is defined as [4]:

\[
d_{sl} (C_i, C_z) = \min \{d(i, j) | i \in C_i, j \in C_z\}
\]

Where \( I \) stands for \( \text{Instance}_i \), and \( j \) stands for \( \text{Instance}_j \).

4.2. Complete Link
The maximum similarity distance between process instances in \( C_i \) and \( C_z \) is selected as the similarity distance between clusters, which is defined as [4]:

\[
d_{cl} (C_i, C_z) = \max \{d(i, j) | i \in C_i, j \in C_z\}
\]

4.3. Group Average
Select the average similarity distance between all process instances in \( C_i \) and \( C_z \) as the similarity distance between clusters, and define as:

\[
d_{ga} (C_i, C_z) = \frac{\sum_{i \in C_i} \sum_{j \in C_z} d(i, j)}{n_i n_z}
\]

Where \( n_i = |C_i| \), \( n_z = |C_z| \) respectively represent the number of process instances in \( C_i \), \( C_z \).

5. Evaluation of cluster results
Silhouette Coefficient is a measure used to express the degree of cohesion and separation of clusters. The coefficient is based on the difference between the average distance of the points of the nearest other cluster and the average distance of the points of the same cluster. For each \( \text{Instance}_i \), \( i \in C_i \), its contour coefficient \( S_i \) is defined as [5]:

\[
s_i = \frac{\mu_{\text{min}}(i) - \mu_{a}(i)}{\max \{\mu_{\text{min}}(i), \mu_{a}(i)\}}
\]

Where \( \mu_{\text{min}}(i) \) is the average distance between all process instances in other cluster \( C_z (z \in [1, k], z \neq t) \):

\[
\mu_{\text{min}}(i) = \min_{z \neq t} \left[ \frac{\sum_{j \in C_z} d(i, j)}{n_z} \right]
\]

And \( \mu_{a}(i) \) is the mean distance between \( \text{Instance}_i \) and all process instances of the same cluster \( C_t \):

\[
\mu_{a}(i) = \frac{\sum_{j \in C_t, j \neq i} d(i, j)}{n_t - 1}
\]
The value range of $S_i$ of each instance is $[-1,1]$. The closer $S_i$ is to 1, the closer the instance is to process instances in the same cluster, and the farther it is from process instances in other clusters, the more reasonable the clustering results are. The closer $S_i$ is to -1, the more the instance should be divided into another cluster and the clustering result is unreasonable. When $S_i$ value is close to 0, it indicates that the instance is on the boundary of two clusters.

After the $S_i$ value of each instance is calculated, the average value of $S_i$ of all instances is taken as the contour coefficient of the entire clustering result, which is defined as:

$$SC = \frac{1}{N} \sum_{i=1}^{N} S_i$$  \hspace{1cm} (17)$$

Meanwhile, the closer the SC value is to 1, the better the clustering result will be.

6. Experimental verification and result analysis

6.1. The validity of clustering algorithm is verified.

According to the relevant data, it can be roughly divided into 6 to 10 clusters from the aspects of subsection division, technological process. In order to avoid the influence of subjective factors of managers, the selection range of clusters was defined as $[5,11]$, and the redundancy of one cluster was left before and after each cluster. The detailed clustering results are shown in FIG. 2 to FIG.5.

![Fig. 2. Clustering results of Euclidean distance](image1)

![Fig. 3. Hamming distance clustering results](image2)
As shown in FIG. 2, when Euclidean distance is used as the similarity distance calculation method between instances in the trajectory clustering process, the similarity distance selection method between the average clusters of the groups is selected. When the number of clusters is set to 11, the clustering result is the best, and the SC value is 0.451. As shown in FIG. 3, when Hamming distance is used as the similarity distance calculation method between instances in the trajectory clustering process, the similarity distance selection method between the average clusters of the groups is selected. When the number of clusters is set to 10, the clustering result is the best, and the SC value is 0.472. As shown in FIG. 4, when using The Jkard distance as the similarity distance calculation method between instances in the trajectory clustering process, the similarity distance selection method between the average clusters of the groups was selected. When the number of clusters was set to 11, the clustering result was the best, and the SC value was 0.43. As shown in FIG. 5, when cosine distance is used as the similarity distance calculation method between instances in the trajectory clustering process, the similarity distance selection method between the average clusters of the groups is selected. When the number of clusters is set as 11, the best clustering result is obtained, and the SC value is 0.428.

At the same time, it can be seen that, no matter use what kind of calculation method of the similarity distance between process instances, using a single cluster similarity distance between selection method to get the clustering result is the worst, the rest of the two methods in clustering similar effect on the performance, when the number of clusters is larger, set the average cluster similarity distance between selection method is superior to the complete chain. In addition, it can also be found that with the increase
of the number of clusters, the overall trend of clustering result score is higher and higher. This is because the more clusters are divided, the similarity between process instances in each cluster is bound to be higher, but it is not absolute. In FIG. 2 to FIG. 5, there are many local optimal advantages. Therefore, when seeking the optimal clustering result within a certain range of the number of clusters, it is not necessarily the more the better.

6.2. Analysis of clustering results

By analyzing the original data of the event log and combining with the segmentation of the case ship, it is concluded that: most of the 17 sections in C1 are double-bottom sections, only one is bottom section of the bow. All the 10 sections in C2 are required by PSPC (special coating). In the field operation and logistics process of the 30 sections in C3, all of them include three tasks of "Inspect 1", "Inspect 2" and "Inspect 3", and basically all of them include three tasks of "Wait 1", "Wait 2" and "Wait 3". The 24 sections in C4 are similar to those in C3, but the task of "Inspect 2" is not included in the logistics process of field operation. The 11 sections in C5 are basically all main deck sections, which have simple structure and do not require pre-outfitting in the yard. All the 6 sections in C6 are fore and aft sections with complex structure and high requirements for construction technology.

Table 1. Optimal clustering results

| Cluster | No. of instances | No. of events | No. of Tasks |
|---------|------------------|---------------|--------------|
| C1      | 17               | 153           | 9e^3         |
| C2      | 10               | 125           | 11e^3        |
| C3      | 30               | 153           | 9e^3         |
| C4      | 24               | 253           | 11e^3        |
| C5      | 11               | 99            | 9e^3         |
| C6      | 6                | 82            | 12e^3        |

7. Conclusion

Aiming at the clustering problem of segmented outfield work logistics process, this paper proposes the method of trajectory clustering, and selects the agglomerative hierarchical clustering algorithm as the basic clustering algorithm of trajectory clustering. Four calculation methods of similarity distance between process instances, such as Euclidean distance and Hamming distance, are introduced. Three similarity distance selection methods of single chain, complete chain and group average are introduced, and the cluster result evaluation method based on contour coefficient is introduced. Finally, the event log of the actual process of segmented outfield logistics is verified by experiments, and the optimal
clustering result is obtained, which verifies the effectiveness of the trajectory clustering algorithm in solving the clustering problem of segmented outfield logistics process.

References
[1] Jeong Y K, Kim B S, Shin J G, et al. A ship block logistics support system based on the shipyard simulation framework[C]. Winter Simulation Conference. IEEE, 2017.
[2] Jeong Y K, Lee P, Woo J H. Shipyard block logistics simulation using process-centric discrete event simulation method[J]. Journal of Ship Production and Design, 2018, 34(2):168-179.
[3] Park C, Seo J, Kim J, et al. Assembly block storage location assignment at a shipyard: a case of Hyundai Heavy Industries[J]. Production Planning and Control, 2007, 18(3):180-189.
[4] Park C, Seo J. Assembly block storage location assignment problem: revisited[J]. Production Planning and Control, 2009, 20(3):216-226.
[5] Park C, Seo J. Mathematical modeling and solving procedure of the planar storage location assignment problem[J]. Computers and Industrial Engineering, 2009, 57(3):1062-1071.