Identification of Class number of ship radiated noise based on RK-Means algorithm

Yuanming Ding\textsuperscript{1,a*}, Lijia Liu\textsuperscript{2,b} and Yang Yang\textsuperscript{3,c}

\textsuperscript{1,2,3} Communication and Network Laboratory Dalian University, Liaoning, China
\textsuperscript{a*} dingyuanming@dlu.edu.cn, \textsuperscript{b} 1152160864@qq.com, \textsuperscript{c} yang_dldx@163.com

Abstract: Ship radiated noise recognition is one of the basic problems in underwater target recognition. To identify the number of ship types according to ship noise, the ship scale can be predicted in advance and the next step of processing can be carried out. Aiming at the problem of identifying the number of ship types based on ship radiated noise, a RK-Means (Recurrent K-Means) is proposed in this paper. According to the short-time energy characteristics of the target signal, the dimension of the feature data is reduced by principal component analysis; The K-Means algorithm is used to cluster and identify the dimension-reduced signals through cyclic iteration; A single connection algorithm is used to determine the appropriate clustering center, which reduces the similarity value of data objects between classes and makes the difference value between classes large. Furthermore, the algorithm can automatically determine the K value, which is the number of ship types. Verified by the experimental data, compared with the contrast algorithm, the K-value recognition result and clustering effect of the RK-Means clustering algorithm are better.

1. Introduction
With the development of marine science and technology, China has carried out in-depth research on the identification and classification technology of ship radiated noise. In order to quickly judge the number of ship types, the methods of feature extraction and clustering recognition can be used. Clustering is a method of data mining, which divides the data into multiple clusters through a certain process, so that the similarity between objects in the cluster is high and the similarity outside the cluster is low\cite{1}. Clustering methods include hierarchical clustering, partition clustering, grid-based clustering and model-based clustering\cite{2}. It is proposed by J M Queen\cite{3} that K-Means algorithm is a wide range of clustering algorithm\cite{4}. Compared with other clustering algorithms, K-Means algorithm is easy to understand and can still ensure good scalability for datasets with large sample size, but the original K-Means algorithm has some limitations. K value needs to be determined artificially, it is sensitive to the selection of initial points, and its adaptability is poor\cite{5}. This will bring some errors to clustering. In view of the shortcomings of the K-Means algorithm, Z T Gao\cite{6} proposed the IBiK-Means algorithm, which makes use of the variation law of the distance between the centroids of each split in the process of the binary K-Means algorithm. D Arthur\cite{7} proposed the K-Means++ algorithm, which randomly selects samples from the data set as the initial clustering center, and each sample has a length, so that the larger the length of the sample is selected as the next initial centroid. The above algorithms solve the problems of automatic setting of K value and local minimum, but the algorithm has too much computation and slow efficiency. When clustering is scattered, K value is easy to be misjudged. In this work, in view of the problems existing in the above research, this paper proposes the RK-Means clustering algorithm for clustering recognition through cyclic iteration, and uses the single connection
algorithm to calculate the difference between clusters to determine the optimal K value, which overcomes the difficulty of manually setting the K value of the original K-Means algorithm, and solves the clustering effect and overcomes the problem of local minimum.

2. RK-MEANS CLUSTERING ALGORITHM

2.1. K-Means Clustering Algorithm

K-Means clustering algorithm is for a large number of untagged data, according to the similarity of each data, this large amount of data is similarly divided into a cluster, each cluster has at least one data object, so that the points in the cluster are connected as closely as possible, each classification will carry on the next iteration, until K clusters classification results are reasonable, belong to unsupervised learning.

In this paper, SSE(Sum of the squared Error) is used as the objective function of clustering. Multiple different clusters generated by running K-Means for many times, the closer the SSE is to 0, the better the model fitting is. The formula is as follows:

\[ SSE = \sum_{i=1}^{k} \sum_{x \in C_i} \text{dist}(C_i, x)^2 \]

In the formula, \( K \) and \( C_i \) denote the number of cluster centers and the \( i \)th center, respectively, and \( \text{dist} \) is the Euclidean distance. The specific flow chart of K-Means is as follows in Figure. 1.

![Figure. 1 flow chart of K-Means algorithm](image)

2.2. RK-Means Clustering Algorithm

The purpose of RK-Means algorithm is to select the optimal K value through exhaustive method and the relationship and distance between clusters to achieve better clustering results. After clustering with each cluster center as the core in each cycle, we can get the cluster set \( S = \{s_1, s_2, \ldots\} \). Here we use the single connection algorithm to calculate the distance between the clusters. The single connection algorithm is that the clustering process terminates when the distance between the nearest clusters exceeds the maximum threshold. It starts with a proximity matrix. In general, given \( n \) objects to be clustered, such as \( X = \{x_1, x_2, \ldots\} \), and the proximity matrix is a symmetric matrix of \( n \times n \), in which the elements on the diagonal of the matrix are zero. Only \( n(n-1)/2 \) elements on the right side of the diagonal are considered. The threshold graph method needs to be considered before using the single connection algorithm. The threshold graph is an undirected graph with two nodes, and each node represents an object. A threshold graph is represented by \( G(v) \), where \( v \) represents a dissimilarity level. Given a \( v \), if the dissimilarity between nodes \( i \) and \( j \) is less than \( v \), an edge is inserted between nodes \( i \) and \( j \).

\[ (i, j) \in G(v), d(i, j) \leq v \]

The distance formula between single connection algorithm clusters is as follows:

\[ d_{\text{min}} = (C_i, C_j) = \min_{p_i \in C_i, p_j \in C_j} |p_i - p_j| \]

Among them, \( |p_i - p_j| \) is the distance between two clusters \( C_i \) and objects \( p_1 \) and \( p_2 \) in \( C_j \).
The distance measurement of single connection algorithm is explained intuitively as shown in Figure 2.

![Figure 2 distance metric diagram of single connection algorithm](image)

Figure 2 distance metric diagram of single connection algorithm

The single connection algorithm uses the Euclidean distance formula to calculate the similarity between each cluster. The smaller the distance is, the higher the similarity is. The nearest distance of the objects in each cluster is combined into a cluster. The algorithm confirms the similarity between the clusters by calculating the distance between each kind of data points and the nearest data points in all clusters. The larger the sum of these shortest distance values is, the smaller the similarity between the clusters is, the K value obtained is the optimal value. The schematic diagram of the algorithm is shown in Figure 3.

![Figure 3 schematic diagram of single connection algorithm](image)

Figure 3 schematic diagram of single connection algorithm

According to the reference [8], there are 7 kinds of civil ships, so the exhaustion number R can be determined. The flow chart of the algorithm is shown in Figure 4.

![Figure 4 flow chart of RK-Means algorithm](image)

Figure 4 flow chart of RK-Means algorithm

3. Data processing of ship radiated noise

In order to better identify the number of ship types, the ship noise data should be processed first. The noise data processing flow is shown in Figure 5.

![Figure 5 data processing flow chart](image)

Figure 5 data processing flow chart
3.1. Feature Extraction

Because the ship radiated noise is a kind of non-stationary random signal, and the data sample is audio data, so the speech analysis is generally short-term analysis. In this paper, the short-time energy is used to extract the feature of the audio signal in the time domain analysis, and the short-time energy reflects the strength of the signal at different times. Assuming that it is the \( m \)-th frame, the short-term energy formula of the audio signal is as follows:

\[
E(m) = \sum_{n=0}^{N-1} X_n^2(n)
\]  

(4)

Among them, \( X_n \) is the speech signal and \( N \) is the frame length of the signal.

3.2. PCA Dimensionality Reduction

Because the short-term energy data belongs to multi-dimensional data, the principal component analysis PCA (Principal Component Analysis) method is used to reduce the dimension. PCA is a method that can reduce high-dimensional data to low-dimensional data. It transforms the original data into a set of linear independent \( x \) representations of various dimensions through linear transformation, which is used to extract the main feature components of the data. The covariance matrix of \( A_a \) and \( A_b \) in PCA is:

\[
\text{cov}(A_a, A_b) = \frac{1}{n-1} \sum_{k=1}^{n} (A_a^{(k)} - \bar{A}_a)(A_b^{(k)} - \bar{A}_b)
\]

(5)

Among them, \( A_a^{(k)} \), \( A_b^{(k)} \) represents the value of the \( k \)-th sample of the feature \( A_a \), \( A_b \), and \( \bar{A}_a \), \( \bar{A}_b \) represent the sample mean.

3.3. Clustering Effect Evaluation Index

In this paper, two effectiveness evaluation indexes in relative measurement are used as evaluation criteria, which are as follows:

1. Segmentation coefficient \( PC \):

\[
PC = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} p_{ij}
\]

(6)

Among them, \( n \) is the size of the data set \( x_i \), the number of samples is arbitrary, \( c \) is the number of categories, and \( p_{ij} \) is the membership degree of category \( j \). The range of \( PC \) values is \([1/k,1]\). The closer the value is to 1, the clearer the division result is and the better the clustering effect is.

2. Partition entropy coefficient \( PE \):

\[
PE = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} p_{ij} \log(p_{ij})
\]

(7)

Among them, the value range of \( PE \) is \([0, \log a k]\). The closer the value is to 0, the better the clustering effect is.

4. Experimental verification and analysis

In this experiment, the noise of six kinds of civil ships, engineering vessels, fishing vessels, harbor vessels, marine development vessels and boats provided in reference [11] [12] are used as the test data of this algorithm. Each kind of ship noise has its own different characteristics, among which each kind of sample has 15 audio data, which are noise generated in different environments. A total of 90 samples are extracted for short-term energy and reduced by PCA, and the RK-Means algorithm is used for seven iterations. Then random ships are selected for seven iterations to verify the 6-2 kinds of ship radiated noise. The algorithm proves that the greater the difference value is, the more \( K \) value can be determined. As shown in Figure. 6.

As you can see from Figure.6, The algorithm can only judge more than two kinds of ships, and cannot identify single kinds of ships. Next, four data sets are tested and compared with IBiK-Means algorithm, K-Means++ algorithm and RK-Means algorithm. When there are 6-2 kinds of ships, 6 kinds of civil ships are randomly allocated. Table 1 shows the results of \( K \)-value recognition accuracy of the three algorithms on the experimental data set.
Figure. 6 differences between clusters after RK-Means iteration 7 times

Table 1 recognition accuracy of the three algorithms in the K value of the data set

| Number of types | IBiK-Means | K-Means++ | RK-Means |
|-----------------|------------|-----------|----------|
| 6               | 100.0%     | 100.0%    | 100.0%   |
| 5               | 83.3%      | 66.7%     | 100.0%   |
| 4               | 66.7%      | 60.0%     | 86.7%    |
| 3               | 80.0%      | 70.0%     | 90.0%    |
| 2               | 86.7%      | 80.0%     | 93.3%    |

It can be seen from Table 1 that when the number of types is 6, the recognition rate of the three algorithms is 100%. When the number of types is 5-2, the RK-Means algorithm can accurately identify the K value, but the recognition rate of the IBiK-Means and K-Means++ algorithm is slightly lower than that of the RK-Means algorithm, indicating the effectiveness of the RK-Means algorithm to identify the K value. In order to compare the clustering performance of the three algorithms in different data sets, select the data set in which the K value can be determined accurately. Table 2 shows the comparison chart of 6-2 clustering results.

Table 2 clustering result indicators

| algorithm    | 6   | 5   | 4   | 3   | 2   |
|--------------|-----|-----|-----|-----|-----|
|              | PC  | PE  | PC  | PE  | PC  | PE  | PC  | PE  | PC  | PE  |
| IBiK-Means   | 0.88| 0.29| 0.79| 0.32| 0.65| 0.43| 0.79| 0.33| 0.72| 0.31|
| K-Means++    | 0.79| 0.31| 0.77| 0.45| 0.62| 0.45| 0.70| 0.36| 0.69| 0.33|
| RK-Means     | 0.91| 0.16| 0.83| 0.21| 0.85| 0.43| 0.88| 0.19| 0.81| 0.20|

As can be seen from Table 2, the performance index of RK-Means algorithm is better than that of IBiK-Means and K-means++. In this paper, the K value is judged according to the difference between clusters calculated by single connection algorithm, which makes the distance measurement and clustering effect between sample points more accurate. In order to quickly judge the number of ship types, the clustering time consumption of this algorithm is compared with that of two algorithms, as shown in Figure 7.

As can be seen from Figure 7, the algorithm in this paper takes the least time, because the algorithm uses exhaustive method to specify the number of iterations, and when the type does not exceed the number of iterations, the time is the same, while IBiK-Means needs to calculate the change of the distance between the centroids of each split. When the number of types is small, the time-consuming is short, while K-Means++ needs to calculate the longest distance of each sample, which takes the most time. Therefore, the algorithm proposed in this paper meets the needs of quickly judging the number of ship types.
5. Conclusion

In this paper, the feature of six kinds of ship radiated noise audio data is extracted, the dimension is reduced by PCA, and the RK-Means algorithm is proposed. In the clustering process, the single connection algorithm is used to calculate the difference between clusters, and the distance is calculated by Euclidean distance, which solves the problems of clustering center selection and clustering effect of the original K-Means algorithm. The experimental data show that the RK-Means clustering algorithm is more accurate than IBiK-Means and K-Means++ algorithm. The algorithm proposed in this paper meets the need of quickly judging the number of types of ships.

Acknowledgment

This work was supported in part by National Natural Science Foundation(61901079).

References

[1] Wang,J.,Wang,S.T.,Deng,Z.H. (2012) A New Text Clustering Method Based on Feature Weighted Distance and Soft Subspace Learning. J. Chinese journal of computers., 35:1655-1665.
[2] Sun,J.G.,Liu,J.,Zhao,L.Y. (2008) Research on Clustering Algorithm. J. Journal of Software., 19:48-61.
[3] Macqueen,J. (1967) Some methods for classification and analysis of multivariate observations. In: Proc of Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: pp: 281-297.
[4] Jain,A.K. (2010) Data clustering: 50 years beyond K-means. J. Pattern Recognition Letters., 31: 651-666.
[5] Xue,Y.X.,Xu,H.W.,Li,L. (2018) Global optimization K-means clustering algorithm based on sample density. J. Computer engineering and application., 54: 143-147.
[6] G.Z.T. (2014) Automatic determination of k in K-Means algorithm. J. Modern computer., 32: 23-27.
[7] Arthur.,David.,S., Vassilvitskii. (2007) K-means++: the advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. pp: 1027-1035.
[8] Deng,D.H. (2015) Classification of civil ships. https://wenku.baidu.com/view/
[9] Bezdek,J.C. (1973) Cluster validity with fuzzy sets. J. Cybernetics and Systems., 3: 58-73.
[10] Bi,K.,Wang,X.D.,Xing,Y.Q. (2015) Fuzzy Clustering Integration Method Based on Fuzzy Measure and Evidence Theory. J. Control and decision-making., 30: 823-830.
[11] Gao,T.L. (2011) Sound effect database of ship noise. https://jianchuan-zaoxshengyinxiao.com/
[12] Dsantos Spain. (2016) ShipsEar: an underwater vessel noise database. http://atlanttic.uvigo.es/underwaternoise/