Improvement of *k*-means Clustering Algorithm for Analyzing the Morphology of Ice Ridge Sails

[1] Bing Tan [2] Zhijun Li [3] Peng Lu [4] Christian Haas [5] Enmin Feng

[*1, First Author] *State Key Laboratory of Coastal and Offshore Engineering, Dalian University of Technology, Dalian* 116024, *China*; *School of Mathematics and Statistics, Nanyang Normal College, Henan 473061,* tanbing111@126.com

[2, Corresponding Author] *State Key Laboratory of Coastal and Offshore Engineering, Dalian University of Technology, Dalian* 116024, *China*, lizhijun@dlut.edu.cn

[3, 4, 5]*State Key Laboratory of Coastal and Offshore Engineering, Dalian University of Technology, Dalian* 116024, *China*, lupeng@dlut.edu.cn, *Department of Earth and Atmospheric Sciences, University of Alberta, Edmonton, T*6*G* 2*E*3, *Canada*, christian.haas@ualberta.ca, *School of mathematical science, Dalian University of Technology, Dalian* 116024, *China,* emfeng@dlut.edu.cn

Abstract

An improved k-means clustering algorithm is proposed after analyzing the disadvantages of the traditional k-means algorithm. The cluster centers are initialized by combining the sample mean and standard deviation, the optimal cluster centers are searched by the hybridizing particle swarm optimization and traditional k-means algorithm, and the criterion function is improved during the iteration process to search the optimal number of clusters. The theory analysis and experimental results show that the improved algorithm not only avoids the local optima, also has greater searching capability than the traditional algorithm. This improved algorithm is used to analyze the morphology of the ridge sail (the upper surface of ice ridges). The comparison with the measured data shows that the influences of the geographical locations and the growing environments on the formation of ice ridges can be perfectly reflected by the clustered results.

Keywords: *Improved k-means clustering, ridge sails, morphology*

1. Introduction

The formation of ice ridges is one of the results on the upper and lower ice surfaces owing to the crushing and piling up of ice blocks [1]. These ridges play key effects on the estimations of the ice mass, thickness, and the momentum and thermal exchange between the atmosphere and sea ice as well as the ocean and sea ice [2]. Significant morphological diversities of ice ridges are caused mainly by the geographical locations and the growing environments, thus the ridges are generally researched by the classifying methods [3, 4]. And the accuracy of the classification has been more and more important.

Cluster analysis is a method of reassigning the elements of a sample set into different clusters according to their similarity. The samples in the same cluster should be similar to each other as much as possible, but samples in different clusters are very dissimilar. The *k*-means clustering is a typical

partitioning method which has been widely applied in data mining and knowledge discovery field [5-7]. Although it is simple, fast, relatively scalable (for large data set) and efficient, traditional *k*-means clustering algorithm also suffers some well-known drawbacks: the number of clusters should be previously fixed; the randomly initialed cluster centers may lead to different results, even the nonexistence of the solution; the extreme value is obtained by Gradient Method (along the decreasing direction of energies), which often leads to the local optimum. These drawbacks significantly limit the application of the traditional *k*-means algorithm.

To overcome the above drawbacks, Genetic Algorithm (GA) has been used to improve the traditional *k*-means clustering algorithm in resent years, and have achieved certain results [8]. Although these algorithms theoretically converge to the global optimum with the probability of 1, the convergence is ensured by the inter-operability of the probability transfer matrix during the variation process, and excessive iterations and the lower clustering accuracy, even fluctuation phenomenon in the late evolution will be led to by the occurrence of the degradation during the evolution process. Particle swarm optimization (PSO) [9] originates from the simulation of the migration and cluster behavior of birds during their feeding process, and it has not only the global searching capability as GA, also strong local searching capability by adjusting parameters. The adjustment of parameters of PSO is simple, and thus more suitable for computer programming and processing. It converges faster to the optimal solution than GA in most cases, and can avoid the regression aroused by complete random searching.

This study proposes an improved *k*-means algorithm based on (PSO), and compares it with the traditional *k*-means algorithm. The improved algorithm is applied in the research of the surface morphology of ice ridge sails, and the cluster results are analyzed by combining the geographical locations and the growing environments of ice ridges finally.

2. Traditional *k*-means clustering algorithm and PSO

2.1. Traditional *k*-means clustering algorithm

Let $\Omega=(x_1, x_2, \cdots, x_n)$ be the sample set, and k the number of clusters, where x_i is a D-dimensional vector. Set $C=(C_1, C_2, \cdots, C_k)$ is a division of the set Ω, which satisfies: $\Omega=\cap_{1\le j\le k} C_j$, $C_j\neq\Phi$ $(1\le j\le k)$, and $C_j \cup C_{j'}=\Phi$, $j\neq j'$, $1\le j, j'\le k$. Then the clustering problem can be expressed as the following.

Definition 1. Define a mapping $f: \Omega \rightarrow C$, such that the *ith* sample x_i of the set Ω is mapped into the *jth* cluster C_j of the division C, where $C_j=\{ x_i| f(x_i)=c_j, x_i \in\Omega, i=1,2,\cdots, n\}$, c_j is the cluster center of C_j, $j=1, 2, \cdots, k$.

In the traditional algorithm, the number of clusters is fixed and the cluster centers are randomly initialized firstly, then the remaining samples of the set are assigned to the nearest cluster by the minimum distance principle. The sample mean of each cluster are taken as the new cluster center, and all the samples are reassigned to the nearest cluster. The process is circled until the criterion function converges. The typical minimum distance principle is

$$d_{ij} = \min_{1\le j\le k} \| x_i - c_j \| \tag{1}$$

Where $\|\cdot\|$ denotes the Euclidean (L_2) norm, and d_{ij} the minimum distance between the sample x_i and

the center c_j , j=1, 2, ⋯, k. Namely, the sample x_i is assigned into the nearest cluster.

The most commonly used criterion function is the squared-error criterion which is defined as

$$J(x_i^{(j)}, c_j; k) = \sum_{j=1}^{k} \sum_{x_i^{(j)} \in C_j} \| x_i^{(j)} - c_j \|^2 \quad (2)$$

where $J(x_i^{(j)}, c_j; k)$ is the square-error sum for all samples in the set Ω.

The main process of the traditional *k*-means clustering algorithm consists of the following steps.

(Step 1.) *k* samples are selected randomly as the initial centers of the *k* clusters.

(Step 2.) Each object x_i (i=1, 2,⋯, n) in the set Ω is assigned into the nearest cluster C_j by the minimum distance principle (1).

(Step 3.) Calculate the sample mean of each cluster: $c_j = 1/ n_j \cdot \sum_{1 \le i \le nj} x_i^{(j)}$, where $x_i^{(j)}$ and n_j are the sample and the number in C_j, respectively (j=1, 2, ⋯, k), and then taken as the new center.

(Step 4.) Criterion (2) is calculated. If $\partial J(\cdot, c_j; k)/\partial c_j = 0$, j=1, 2, ⋯, k, stop; else go to (Step 2).

From the above process, we can see that the results of the traditional *k*-means clustering algorithm are very sensitive to the initial centers, and also impacted the number of clusters: the randomly selected initial cluster centers may lead to unreasonable results, and an inappropriate number *k* may lead to unreasonable clusters which can't represent certain characteristics of the sample set. Additionally, the convergence of the criterion function is judged by its gradient, which easily leads the algorithm to the local optimum.

2.2. Particle swarm optimization (PSO)

PSO originates from the simulation of the migration and clustering behavior of birds during their feeding process. By using fully the intelligence of the group and their own, the individuals in the group search for the optimal region in the complex space by constantly adjusting and learning. It is a class of stochastic global optimization algorithm based on the iteration [9]. In PSO, each solution is taken as a "particle" in the search space, and flights to the better position according to its own "experience" (the optimal solution searched by itself, e.g., the individual optimal position $p_{best\text{-}s}$ with the fitness value p_{best}) and the optimal "experience" of the group (the optimal solution searched by the group so far, e.g., the global optimal position $g_{best\text{-}s}$ with the fitness value g_{best}), until the optimal solution is obtained (the advantage of a solution is judged by a fitness function).

Let N be the total number of particles in the group, s_i=(s_{i1}, s_{i2}, ⋯, s_{iD})^T, v_i=(v_{i1}, v_{i2}, ⋯, v_{iD})^T, and $p_{best\text{-}s}$=(p_{i1}, p_{i2}, ⋯, p_{iD})^T the position, velocity, and individual extreme value of the *ith* particle respectively, i=1, 2, ⋯, N, and $g_{best\text{-}s}$=(p_{g1}, p_{g2}, ⋯, p_{gD})^T the global extreme value. The iteration equations of the *(t+1)th* generation in the standard PSO are then

$$v_{id}^{t+1} = w v_{id}^{t} + c_1 r_1 (p_{id}^{t} - s_{id}^{t}) + c_2 r_2 (p_{gd}^{t} - s_{id}^{t}) \quad (3)$$

$$s_{id}^{t+1} = s_{id}^{t} + v_{id}^{t+1}, \quad 1 \le i \le N, 1 \le d \le D \quad (4)$$

where w is the inertia weight, the value of which determines the degree of inheritance of the current velocity of the particle. c_1 and c_2 are learning factors, and usually taken as c_1=c_2=2. r_1 and r_2 are random numbers uniformly distributed in the interval (0, 1).

Because the effect of the previous victory of the particle on the current victory is mainly determined by the inertia weight *w*, the adjusting of *w* thus can achieve a balance between the local and global search: the larger value of *w*, the stronger global search ability of the algorithm; while the algorithm tends to local search for the smaller *w*. Therefore, an appropriate value of *w* not only can improve the performance and optimization of the algorithm, but also reduce the times of the iteration. The linear differential decreasing strategy [10] is used to improve the inertia weight in this study

$$dw(t)/dt = 2(w_s - w_e)/T_{\max}^2 \cdot t \tag{5}$$

Namely,

$$w(t) = w_s - (w_s - w_e)/T_{\max}^2 \cdot t^2 \tag{6}$$

where w_s is the inertial inertia weight, w_e is the inertia at the end of the iteration, t is the current iterations, and $T_{\max}$ the total iterations. Generally, w_s=0.9 and w_e =0.4 is used.

3. Improved *k*-means clustering algorithm

3.1. Improved *k*-means clustering algorithm

Because all the samples are clustered from the initial centers, then if better initial cluster centers are selected at the very start, the algorithm should require fewer iterative procedures before converging. Hence, the strategy for selecting the initial cluster centers greatly affects the computational complexity of the *k*-means algorithm. According to the distribution characteristics of the random function, the elements in the sample set distribute mainly near the sample mean. Let η be the mean of the sample set, and σ the standard deviation, then the initial cluster centers can be randomly selected from the interval $[\eta\text{-}\sigma, \eta+\sigma]$.

The optimal cluster centers are searched by the combining of PSO and the traditional *k*-means clustering algorithm. And to obtain the optimal number of clusters, a function related to the squared-error criterion (2) is introduced during the clustering process to guide the updating of *k* [11].

The improved algorithm works as following.

(Step 1.) Let *k* be the number of clusters, *N* the size of the particle group, $T_{\max}$ the total iterations, and set *t*=0. Initialize the particle group *P*(0): select randomly *k* samples from the interval $[\eta\text{-}\sigma, \eta+\sigma]$ as the initial centers, and cluster by the traditional *k*-means clustering algorithm. Circle *N* times, e.g., produce *N* particles (positions). Initialize the velocity of each particle, set the initial position of each particle as the individual optimal position $p_{best\text{-}s}$, and the optimal $p_{best\text{-}s}$ as the global optimal position $g_{best\text{-}s}$.

(Step 2.) Process of PSO.

(2.1) Calculate the fitness of the particle s_i^t: $f_i^t = f(s_i^t) = k/(1+J(\cdot, \cdot; k))$, and initialize the velocity of s_i^t, i=1, 2, ···, N;

(2.2) If $f_i^t > p_{best}$, set $p_{best} = f_i^t$ and $p_{best\text{-}s} = s_i^t$, i=1, 2, ···, N;

(2.3) If there exists at least one f_i^t (i=1, 2, ···, N) such that $f_i^t > g_{best}$, set $g_{best} = f_m^t = \max_{1\le i\le N}\{f_i^t\}$, and $g_{best\text{-}s} = s_m^t$.

(2.4) Calculate the velocity and position of the particle s_i^{t+1} by combing equations (3), (4) and (6),

i=1, 2, ⋯, N, and obtain the (t+1)*th* particle group P(t+1).

(Step 3.) Determine the cluster division of each particle in the group P(t+1) by the minimum distance principle (1). Set $t = t + 1$, if $t < T_{\max}$, go to (Step 2.).

(Step 4.) Output the optimal particle and the corresponding cluster division. Set $Q(k) = J(\cdot, \cdot; k)$.

(Step 5.) For each cluster, calculate the distance between every sample and the cluster center, and calculate the averaged distance: $\bar{d}_j = 1/n_j \cdot \sum_{x_i^{(j)} \in C_j} \| x_i^{(j)} - c_j \|$, j =1, 2, ⋯, k. Select the cluster with the largest average distance and divide it into two clusters, and recalculate $J_j(\cdot, \cdot; k+1) = \sum_{x_i^{(j)} \in C_j} \| x_i^{(j)} - c_j \|^2$, j = 1, 2, ⋯, k + 1. Set $Q(k + 1) = \min_{1 \le j \le k+1} J_j(\cdot, \cdot; k+1)$.

(Step 6.) Remove the cluster with the least number of data in (Step 4), then move all samples of it into the nearest other cluster, recalculate $J_j(\cdot, \cdot; k\text{-}1)$, j = 1, 2, ⋯, k - 1. Set $Q(k\text{-}1) = \min_{1 \le j \le k-1} J_j(\cdot, \cdot; k\text{-}1)$.

(Step 7.) Set k =argmax{ $Q(k\text{-}1)$, $Q(k)$, $Q(k + 1)$}.

(Step 8.) Repeat (Step 2) ~ (Step7) until k remains unchanged.

In the improved algorithm, the initial cluster centers are selected by combining the sample mean and the standard deviation which can not only avoid the occurrence of the unrealistic results, but also reduce greatly the iterations. And the next generation particle group produced by PSO is very random, not only can efficiently overcome the drawback of falling easily into the local minimum, but also converges faster owing to the nonexistence of the degradation.

3.2. Assessment of the cluster results

To assess the results of the clustering algorithms, other two parameters are introduced except for the criterion function (2), e.g. the maximum distance within a cluster ($d_{IC\max}$) and the minimum distance between clusters ($d_{BC\min}$). The distance within a cluster is the average Euclidean distance between a sample and the corresponding cluster center. The maximum distance within a cluster is defined as

$$d_{IC\max} = \max_{1 \le j \le k} \{1/n_j \cdot \sum_{x_i^{\{j\}} \in C_j} \| x_i^{\{j\}} - c_j \|\} = \max_{1 \le j \le k} \bar{d}_j \tag{7}$$

The distance between clusters is the Euclidean distance between any pair of cluster centers. And the minimum distance between clusters is

$$d_{BC\min} = \min_{1 \le j, j' \le k} \{\| c_j - c_{j'} \|\} \tag{8}$$

4. Application in the research of the morphology of ice ridge sails

The morphology parameters of the ice ridge upper surface include mainly the sail height (h), frequency (μ: 1/km), ridging intensity (R_i=<h>/<s>, where <h> is the mean sail height, <s> is the mean sail spacing), sail width, and sail cross-section area. Based on the assumption that all ridges have symmetric triangular cross sections with a similar slope angle, the sail width and cross-section area are w=2$h$cotφ and S=<h>^{2}cotφ, respectively, where φ is the ridge slope angle, and φ=26° is used in this paper as Dierking [3].

Data sets of sea ice surface elevation used in this study were obtained by Alfred Wegener Institute for Polar and Marine Research from August 24 to October 29, 2006, using a helicopter-borne laser altimeter. According to the Rayleigh criterion [3], we extract ridge sails from the height profiles of sea ice surface.

Dierking [3] showed that the ridging intensity R_i should be selected as a quantitative index for the cluster of profiles because changes in the sail height distribution were generally coupled to the changes in the sail spacing distribution. Here we also use the ridging intensity R_i as a cluster index.

The traditional and improved *k*-means clustering algorithm are both employed to cluster the profiles (the number of samples is n=94) in the following. The sample mean is η=0.0169, standard deviation is σ =0.0145. In the improved algorithm, the size of the particle group is N=10, the learning factors are c_1=c_2=2, the initial cluster centers are selected from the interval [0.0024, 0.0314], and the max generation is $T_{\max}$= 500.

The results of the improved algorithm show that k =3 is the optimal number of clusters. The cluster results of the traditional and improved *k*-means algorithm for k =3 are compared in Fig. 1(a, b) (the corresponding clusters are donated by C_1, C_2 and C_3, respectively). The cluster results of the traditional *k*-means algorithm are shown in Fig. 1(a), obviously, the boundaries of the clusters are not clear and difficult to distinguish, and there exist some profiles with different formation mechanism and ridge age in the same cluster. These phenomena don't exist in the results of the improved algorithm (Fig. 1(b)), which indicate better results of the improved *k*-means algorithm than that of the traditional algorithm.

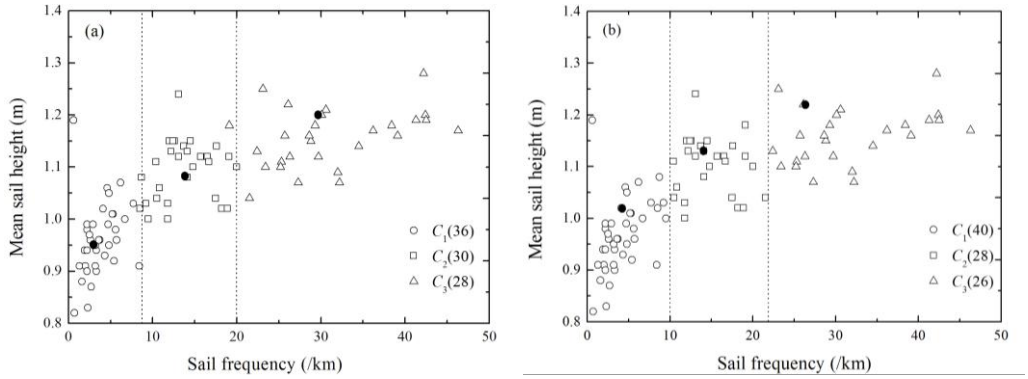

Figure 1. Cluster results of (a) traditional and (b) improved *k*-means algorithms. Optimal centers are donated by solid points. The number of samples for each cluster is given in parentheses

Table 1 is the comparison results between the two algorithms, indicating that the three assessment parameters of the improved algorithm are all better than those of the traditional *k*-means algorithm, and reflecting that the improved *k*-means algorithm is better than the traditional algorithm.

Table 1. Comparison of the traditional and improved *k*-means algorithm

***k-means* Algorithm**	J(×10^4)	$d_{IC\max}$(×10^4)	$d_{BC\min}$(×10^4)	***Centers* (for C_1, C_2 and C_3)**
Traditional	37.1	7.9	114	0.0036, 0.0152 and 0.0361
Improved	34.6	7.7	122	0.0043, 0.0165 and 0.0314

From Fig. 1(b) we can see that the differences of the three clusters are very obvious: the sail frequencies are all smaller than 10/km in the cluster C_1, larger than 10/km, but smaller than 22/km in the middle cluster C_2, whereas all larger than 22/km with the largest 46/km in the upmost cluster C_3. The diversities of the mean sail heights of the three clusters are small (ranging from 0.8 m to 1.3 m). The number of samples in the cluster of C_1 is the largest (about 42.6% to the total), the proportions of other two clusters are 29.7% (C_2) and 27.7% (C_3) to the total samples respectively, representing well statistics significance.

To obtain the better statistical representation, the average ridging intensity, mean sail height, frequency, width and cross-section area for different clusters obtained by the improved algorithm are listed in Table 2. It is obvious that the average sail height increases slowly from 0.96 m to 1.14 m, while the average frequency increases vapidly from 4/km to 32/km with increasing ridging intensity, which indicate a much larger change speed of the average sail frequency than height with the ridging intensity. The differences of mean sail width and cross-section area of the three clusters are not large. The values of the parameters indicate no distinctive variances of the ice ridge surface morphology, e.g., the ridges have similar shapes, although the deformation of sea ice in different regions of the northwestern Weddell Sea varies wildly.

Table 2. Morphology parameters of the clustered profiles obtained by the improved algorithm

Cluster	$\langle R_i \rangle$	$\langle w \rangle$/m	$\langle h \rangle$/m	$\langle \mu \rangle$/(/km)	$\langle S \rangle$/m^2
C_1	0.004	3.88	0.96	4.2	1.88
C_2	0.017	4.40	1.08	15.0	2.42
C_3	0.037	4.64	1.14	31.9	2.69

The comparison results in the measured data shows that the samples in the cluster C_1 occur mainly in the marginal ice zone (MIZ) and Larsen polynyas, and the ridging intensities and frequencies are smaller due to the lower overlapping and rafting rate of floe ice; nearly all samples in the cluster C_2 exist on the band of first- and second-year ice (FYI and SYI) in the center investigated region, the ridging intensities and frequencies are larger relatively due to the dynamic action of FYI or SYI and the refreezing of sea ice in the next winter which didn't melt completely in the summer; samples in the cluster C_3 occur only in the stationary ice pack adjacent to the shelf ice edge of the southern investigation region, and the ridges are formed mainly by the movement of glacial under the dynamic force (such as wind, currents and waves). The above analysis shows that the cluster results of the improved algorithm reflect perfectly the important influence of the geographical locations and the environmental conditions on the formation of the ridges.

5. Conclusions

Theoretical analysis and data experimental results show that the proposed algorithm in this study not only overcomes the drawbacks of traditional *k*-means clustering algorithm, also has a faster convergence rate, and thus is more efficient for clustering analysis. The cluster results of ice ridge sails obtained by the improved algorithm perfectly reflects the important influence of the geographical

locations and environmental conditions on the formation of ice ridges.

Acknowledgments

We are grateful to Dr. Nicolaus Marcel of the German Alfred Wegener Institute for Polar and Marine Research for his help during field investigation. This study was supported by the National Nature Science Foundation of China (No. 40806075), the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (No. 50921001), the Nature Science Foundation of Henan province, China (No. 102300410184), and the special project of Nanyang Normal College, Henan, China.

References

[1] Torge Martin, "Comparison of different ridge formation models of Arctic sea ice with observations from laser profiling", Annals of Glaciology, vol.44, no.1, pp.403-410, 2006.

[2] Tina Tin, Martin O. Jeffries, "Morphology of deformed first-year sea ice features in the Southern Ocean", Cold Regions Science and Technology, vol.36, no.(1-3), pp.141-163, 2003.

[3] Wolfgang Dierking, "Laser profiling of the ice surface topography during the Winter Weddell Gyre Study 1992", Journal of Geophysical Research, vol.100, no.C3, pp.4807-4820, 1995.

[4] Adolphs Ute, "Roughness variability of sea ice and snow cover thickness profiles in the Ross, Amundsen, and Bellingshausen Seas", Journal of Geophysical Research, vol.104, no.C6, pp.13,577-13,591, 1999.

[5] Bing He, Gang Liu, Yuanyuan Wang, "Automatic Image Registration Using Improved LBG Algorithm", Journal of AICIT, vol.6, no.3, pp.257-263, 2011.

[6] Chengjie Gu, Shunyi Zhang, Xiaozhen Xue, "Internet Traffic Classification based on Fuzzy Kernel K-means Clustering", Journal of AICIT, vol.3, no.3, pp.199-209, 2011.

[7] Meng-Dar Shieh, Tsung-Hsing Wang, Chih-Chieh Yang, "A Clustering Approach to Affective Response Dimension Selection for Product Design", Journal of AICIT, vol.6, no.2, pp.197-206, 2011.

[8] Ujjwal Maulik, Sanghamitra Bandyopadhay, "Genetic algorithm-based clustering technique", Pattern Recognition, vol.33, no.9, pp.1455-1465, 2000.

[9] Li Li, Ben Niu. "Particle swarm optimization", Metallurgical Industry Press, China, 2008.

[10] Jianxiu Hu, Jianchao Zeng, "Selection on Inertia Weight of Particle Swarm Optimization", Computer Engineering, vol.33, no.11, pp.193-195, 2007.

[11] Jürgen Beringer, Eyke Hüllermeier, "Online clustering of parallel data stream", Data & Knowledge Engineering, vol.58, no.2, pp.180-204, 2006.