https://doi.org/10.5194/egusphere-2025-2884
Preprint. Discussion started: 14 July 2025

Research on a microseismic signal picking algorithm based on GTOA clustering

Hongjie Zheng[1], Yongqing Wang[1], Qisheng Zhang[1], Zewen Li[1], and Lin Chao[1]

[1]School of Geophysics and Information Technology, China University of Geosciences (Beijing), Beijing, China

Correspondence: Yongqing Wang (yqingwang@yeah.net)

Abstract. Clustering is one of the challenging problems in machine learning. Adopting clustering methods for the picking of microseismic signals has emerged as a new approach. However, due to the low separation of signals generated by microseismic events in real environments and the presence of noise data, existing clustering methods often struggle to achieve satisfactory results. To address these challenges, this paper proposes a clustering method based on the Group Theory Optimization Algorithm (GTOA), combined with the Akaike Information Criterion(AIC) to form an innovative microseismic signal picking algorithm. The GTOA clustering method overcomes the shortcomings of traditional mean clustering algorithms, which are easily influenced by the initial positions of centroids, thus improving the quality of clustering results and enhancing the accuracy of microseismic signal picking.Experiments evaluate the clustering effects using three performance metrics, statistically analyzing the computational results of the GTOA clustering algorithm against six other clustering methods based on evolutionary algorithms across four datasets. The comparative results indicate that the GTOA clustering method outperforms other algorithms in terms of solution quality and exhibits good robustness. Finally, using the GTOA clustering method as the foundation for the designed microseismic signal picking algorithm, experiments comparing it with traditional signal picking methods on low signal-to-noise ratio time series data show that the GTOA-based microseismic signal picking algorithm designed in this paper achieves good picking accuracy.

1 Introduction

Against the backdrop of continuously growing global energy demand, oil and natural gas remain key fossil fuels that have made significant contributions to economic development and improvements in living standards. However, as conventional oil and gas resources gradually deplete, the urgency of developing emerging oil and gas resources has become increasingly prominent, emerging as an important issue that needs to be addressed in the current energy sector (Wang et al., 2021b). In this context, microseismic monitoring technology, as an efficient geological survey tool, has been widely applied in oil and gas exploration. Its capability lies in providing crucial information for the study of deep geological structures and the prediction of resource distribution. By monitoring minor fault activity and fracture propagation underground, microseismic monitoring technology not only provides a scientific basis for the efficient extraction of resources but also offers essential support for engineering safety and environmental protection(CHEN et al., 2020; Jiang et al., 2021; Liu et al., 2021).

https://doi.org/10.5194/egusphere-2025-2884
Preprint. Discussion started: 14 July 2025

Due to the small amplitude of microseismic signals and high environmental noise levels, these signals are often masked by noise, affecting their effective utilization. Noise reduction processing may compromise the fidelity of the signals, subsequently impacting subsequent analyses. The accuracy of microseismic signal detection is crucial for the overall effectiveness of microseismic monitoring. Traditional arrival time picking algorithms, such as Short-Time Average/Long-Time Average(STA/LTA) (Chen et al., 2023) and Akaike Information Criterion(AIC)(Li et al., 2022a), are classic methods widely used in microseismic signal identification. These methods can effectively identify microseismic signals under ideal conditions, but their performance often fails to meet expectations when the signal-to-noise ratio (SNR) decreases. For instance, the STA/LTA method relies on the correlation between signal strength and time, which can severely affect the determination of the validity of the arrival signal in the presence of background noise. Similarly, AIC is prone to noise interference in low SNR environments, potentially leading to misjudgment and hindering the accurate detection of microseismic events (QIU and LI, 2023; Li et al., 2022a). Therefore, the effectiveness of traditional methods is questioned in complex geological environments, highlighting an urgent need for new technological approaches to improve the detection of microseismic signals.

In response to these challenges, many researchers have gradually begun exploring the application of clustering algorithms in microseismic signal picking in recent years. Clustering algorithms can handle large-scale data and significantly enhance the detection capability of microseismic signals by comprehensively analyzing multiple features of the signals. Compared to traditional methods, clustering algorithms demonstrate clear advantages when dealing with low signal-to-noise ratio (SNR) conditions. This approach integrates information from various signal features into a cohesive whole through the fusion of multiple characteristics, effectively reducing the interference from noise and improving the identification capability for microseismic events (Feng et al., 2023; Jin et al., 2024).

Although clustering algorithms have gradually gained attention in microseismic signal picking, related research still faces some shortcomings. Traditional clustering methods are sensitive to the choice of initial values and have poor robustness against the interference of outliers, making it challenging to effectively handle the unique nonlinear characteristics of microseismic signals.

To overcome the issues associated with traditional clustering methods, evolutionary algorithms(EAs) can be employed as optimization engines for the objective function in clustering, providing global search capabilities through intelligent search mechanisms, thereby effectively avoiding local optima. (Ma et al., 2021; Deng et al., 2021; Song et al., 2024) Such clustering algorithms are better suited for handling noisy signals and capturing the nonlinear characteristics of microseismic signals. It is evident that when using clustering approaches based on evolutionary algorithms for microseismic signal picking, the performance of the evolutionary algorithm will directly affect the detection effectiveness of microseismic signals. The Group Theory-based Optimization Algorithm (GTOA) is a novel evolutionary algorithm that introduces mathematical principles from group theory, offering stronger theoretical support compared to biologically inspired evolutionary algorithms (such as GA(Mirjalili and Mirjalili, 2019; Ileberi et al., 2022) and PSO(Gad, 2022; Camacho-Villalón et al., 2021)). It enhances the advantages of evolutionary algorithms while further improving convergence speed and solution accuracy.(Li et al., 2022b; Wang et al., 2021a)

https://doi.org/10.5194/egusphere-2025-2884
Preprint. Discussion started: 14 July 2025

This paper proposes a microseismic signal picking algorithm based on GTOA clustering. The algorithm is founded on the GTOA clustering method, which filters out the sample point index range containing microseismic events and combines it with the AIC algorithm for precise signal picking. When addressing the clustering problem using GTOA, the evolutionary operators are adjusted to enhance the optimization capability of the algorithm by incorporating optimal guiding strategies and adaptive parameter adjustment strategies. Through this algorithm, the accuracy of microseismic signal picking is improved to meet the increasing demands of modern oil and gas exploration and development.

2 Group Theory-based Optimization Algorithm

GTOA is an evolutionary algorithm based on set algebra systems. It promotes cooperation and exploration within the population by defining the interactions and information-sharing mechanisms among individuals. GTOA has advantages such as a simple structure, fewer control parameters, fast convergence speed, and high solution accuracy. This section will first introduce the evolutionary operators in GTOA, followed by a description of the algorithm's pseudocode.

2.1 Evolutionary Operators of GTOA

Let $G_n = \{0, 1, ..., n-1\}$, n is a positive integer and n $\geq$ 2. Define the operation of G_n as Formula 1, then G_n is called a group of operation$\oplus$.

$$a \oplus b = (a+b) mod n \quad a, b \in G_n \tag{1}$$

If G_1, G_2, ..., G_k are k groups, their direct product is denoted as: G$[n_1, n_2, ..., n_k] = G_1 \times G_2 \times \cdots \times G_k$. It can be derived that G$[n_1, n_2, ..., n_k]$ also forms a group. The elements of this direct product group are ordered k-tuples such as $(a_1, a_2, ..., a_k)$, represented as the vector $[a_1, a_2, ..., a_k]$.In $G[m_1+1, m_2+1, ..., m_n+1]$, three distinct random individuals are selected, such as $X = [x_1, x_2, ..., x_n]$, $Y = [y_1, y_2, ..., y_n]$, and $Z = [z_1, z_2, ..., z_n]$. Through the Random Linear Combination Operator (RLCO), algebraic group operations are performed on X, Y, and Z to generate a new individual $R = [r_1, r_2, ..., r_n] \in G[m_1+1, m_2+1, ..., m_n+1]$, achieving group evolution. This approach allows the new individuals to learn from other individuals. It not only enhances the randomness of the generated individuals but also effectively retains characteristics and structures from the original individuals. Therefore, GTOA possesses strong global exploration capabilities, increasing the likelihood of finding high-quality solutions in the solution space. The RLCO formula is shown as 2.

$$R = X \oplus (F(Y \oplus (-Z))) \tag{2}$$

Here, $r_j = x_j \oplus [f_j(y_j \oplus (m_j + 1 - z_j)]$, $j = 1, 2, ..., n.F = [f_1, f_2, ..., f_n] \in \{-1, 0, 1\}^n$ is a vector of random combination factors.

Evolutionary algorithms not only require strong global exploration capabilities but also needs a strong exploitation ability; both are essential for the efficient operation of evolutionary algorithms (He and Wang, 2021).To achieve this, GTOA employs a local search operator based on inverse operations in space G ——the Inversion and Random Mutation Operator (IRMO). The

https://doi.org/10.5194/egusphere-2025-2884
Preprint. Discussion started: 14 July 2025

calculation formula is as follows 3:

$$IRMO(R) = \begin{cases} rand\left([0,m_j]-\{r_j\}\right) & h_1 < P_m, h_2 > 0.5 \\ m_j+1-r_j & h_1 < P_m, h_2 \le 0.5, x_j \ne 0 \quad j=1...n \\ r_j & otherwise \end{cases} \tag{3}$$

P_m is the mutation probability, satisfying $0 < P_m \le 0.5$. h_1 and h_2 are two random numbers uniformly distributed in the interval (0, 1). $rand([0,m_j]-\{r_j\})$ represents a random integer within the range $[0,m_j]$ that is not equal to r_j.

However, IRMO also has its limitations, as it is not suitable for 0-1 vectors in the search space G[2,2,...,2]. To address this, a local search operator suitable for the space G[2,2,. . . ,2] is proposed, drawing inspiration from the mutation operator in genetic algorithms (GA). The calculation method is shown as 4.

$$SMO(R) = \begin{cases} 1-r_j & h < P_m \\ r_j & otherwise \end{cases} \quad j=1...n \tag{4}$$

P_m is the mutation probability, satisfying $0 < P_m \le 1.h \sim U[0,1]$ is a random number.

2.2 The pseudo-code of GTOA

Let n be the dimensions of the objective function, P_m be the mutation probability, N be the population size, T be the maximum number of iteration, and $P(t)$=$\{X_i(t)|1 \le i \le N\}$ be the population of GTOA in iteration t, where $0 \le t \le T$. $X_i(t) = [x_{i1}(t), x_{i2}(t), ..., x_{in}(t)]$ is the i-th individual. The pseudo-code of GTOA is described as 1:

3 GTOA clustering method

3.1 Evolution mechanism adjustment

When applying GTOA for clustering to handle large-scale problems, RLCO faces several challenges, primarily because it lacks information about the global optimal solution, which prevents GTOA from quickly locating potential better solution areas. To address this issue and accelerate the convergence speed, RLCO has been improved by introducing a mechanism based on optimal individual learning. During the evolution process, we use the currently known optimal individual along with two other randomly selected individuals, R and S, from the population to generate new individuals, thereby enhancing the exploration capability and optimization efficiency of the algorithm. The formula for RLCO1 is shown as 5:

$$X = H \bigoplus (R \bigoplus (-S))) \tag{5}$$

Where $H = [h_1, h_2, ..., h_n]$ is the individual with the highest current fitness value, $x_j = h_j \oplus [f_j(r_j \oplus (m_j + 1 - s_j)]$,$j = 1,2,...,n.F = [f_1, f_2, ..., f_n] \in \{-1,0,1\}^n$ is the random combination factor vector.

https://doi.org/10.5194/egusphere-2025-2884
Preprint. Discussion started: 14 July 2025

Algorithm 1 Algorithm of GTOA

Input: The objective function min$f(X)$,Parameters$N$ and T,Mutation probability P_m.
Output: An approximate solution (or optimal solution) H and $f(H)$.

```
Generate initial population P(0) ={X_i(0)|i = 1···N} randomly;
Calculate f(X_i(0)),i = 1···N,t ← 0;
while t < T do
    for i ← 1 to N do
        Y ← X_{p1}(t) ⊕ (F(X_{p2}(t) ⊕ (−X_{p3}(t))));
        Y ← IRMO(Y,P_m) or Y ← SMO(Y,P_m);
        if f(Y)>f(X_i(t) then
            X_i(t+1) ← Y;
        else
            X_i(t+1) ⟵ X_i(t);
        end if
    end for
    t ← t+1;
end while
Determine the best individual H in P(T);
return (H,f(H))
```

The mutation probability (P_m) is a key hyperparameter in swarm intelligence optimization algorithms, and its setting directly affects the global search capability and convergence accuracy of GTOA. Traditional methods typically set P_m as a fixed value, which may lead the algorithm to fall into local optima or insufficient convergence speed in complex optimization problems. To better apply GTOA to clustering algorithms, a dynamic mutation probability strategy based on iterative count adaptation is proposed. By dynamically adjusting P_m using formula 6, its value increases with the number of iterations, thereby better balancing the exploration and exploitation capabilities of the algorithm. In the initial stage of the search, the diversity of the population is relatively high; thus, a smaller mutation probability is assigned to individuals to achieve stable convergence results. In the later stages of the search, a higher mutation probability is given to individuals to enhance local search.

$$P_m = P_{min} + (P_{max} - P_{min}) * t/T \tag{6}$$

Where P_{max} and P_{min} represent the upper and lower bounds of P_m, respectively, t is the current iteration count, and T is the maximum iteration count.

3.2 Representation of the algorithm's solution

In clustering algorithms, the output representation should intuitively display the classification results. Therefore, in GTOA, each individual uses integer vector encoding, where the encoding results directly correspond to a partitioning scheme of the

https://doi.org/10.5194/egusphere-2025-2884
Preprint. Discussion started: 14 July 2025

dataset. Each component of the individual's vector encoding represents the cluster membership label of the corresponding sample, achieving a clustering representation of the entire dataset through this encoding method. For example, the output encoding vector X = [01011] corresponds to partitioning the dataset into 2 clusters, with the first cluster containing samples 1 and 3, and the second cluster containing samples 2, 4, and 5. This encoding method intuitively presents the clustering division results of the dataset.

3.3 Population initialization

In the clustering optimization problem, the generation of the initial population directly affects the convergence speed of the algorithm and the quality of the solutions. In the population initialization phase of GTOA, first, m initial centroids are randomly generated for each individual in the sample space. The selection of these centroids is random to ensure that individuals can cover the diversity of the sample space in the initial stage. Next, the distance between each sample point and these centroids is calculated, and based on the proximity, sample points are assigned to the nearest clusters, with all sample points in the same cluster being assigned the same integer label. Finally, individuals are encoded based on the labels of the sample points, thus establishing a suitable data structure for the subsequent optimization process.

3.4 Individual fitness evaluation

The sum of squared errors (SSE) is used as the fitness evaluation criterion for individuals. SSE has a good ability to measure internal consistency and can effectively gauge the closeness of sample points to the centroids, thereby reflecting the quality of the clustering effect. A smaller SSE value indicates a better clustering outcome. The computation formula for SSE is shown as 7:

$$SSE = \sum_{k=1}^{m} \sum_{i=1}^{|c_k|} \|p_i - c_k\|^2 \tag{7}$$

Here,$\|\cdot\|$ represents the Euclidean distance,$|c_k|$ denotes the number of samples contained in the cluster with centroid c_k, and p_i indicates the i-th sample point in that cluster.

In each iteration, according to the coding of the newly evolved individuals, samples in the dataset with the same label are grouped into the same cluster. For each cluster, the mean of all sample points contained within it is calculated as the centroid position of that cluster. An array $C = \begin{bmatrix} c_{11}, & \cdots, & c_{1d}, & \cdots, & c_{m1}, & \cdots, & c_{mn} \end{bmatrix}$, where $c_{k1},\cdots,c_{kn}$ represents the coordinate positions of the k-th centroid c_k, is used to store the centroid coordinates. After obtaining the new centroid positions, the distances from the sample points within each cluster to the centroid are calculated. Finally, the sum of squared errors (SSE) is used as the evaluation index, which represents the current fitness value of the individual.The relevant pseudocode description is shown as 2.

https://doi.org/10.5194/egusphere-2025-2884
Preprint. Discussion started: 14 July 2025

Algorithm 2 Algorithm of SSE

Input: Number of categories m, Number of features d , Data set $\{p_j \mid j = 1\ldots n\}$,$X = [x_1, x_2, ..., x_n]$,$C = [0, 0, ..., 0]$.
Output: Centroids array C, Fitness value(SSE)

```
SSE ← 0;
for k ← 1 to m do
    count ← 0;
    for j ← 1 to n do
        if x_j = k then
            count ← count + 1;
            for f ← 1 to d do
                c_kf ← c_kf + p_jf
                SSE ← SSE + (p_jf − c_kf)^2
            end for
        end if
    end for
    if count > 0 then
        for f ← 1 to d do
            C_kf ← C_kf / count;
        end for
    end if
end for
return (C, SSE)
```

3.5 Overall process of the algorithm

GTOA clustering method is shown as figure 1: First, the input dataset is normalized and a population containing N individuals is randomly initialized, with each individual consisting of m random centroids. The algorithm evaluates the clustering quality by calculating the sum of squared errors (SSE) for each individual and records the optimal individual H. During the iterative optimization phase, the algorithm uses a group teaching mechanism (RLCOI) to generate new solutions and selects either Improved Random Mutation (IRMO) or Simple Mutation (SMO) for mutation operations based on the number of clusters m, in order to enhance population diversity. After each iteration, the centroids of the new solutions are recalculated and their SSE is evaluated, retaining the better solutions. When the maximum number of iterations T is reached, the algorithm outputs the optimal centroid set H with the minimum SSE, completing the clustering process. This algorithm effectively balances global exploration and local development capabilities through improved evolutionary mechanisms and adaptive mutation strategies, making it suitable for clustering tasks with complex datasets.

https://doi.org/10.5194/egusphere-2025-2884
Preprint. Discussion started: 14 July 2025

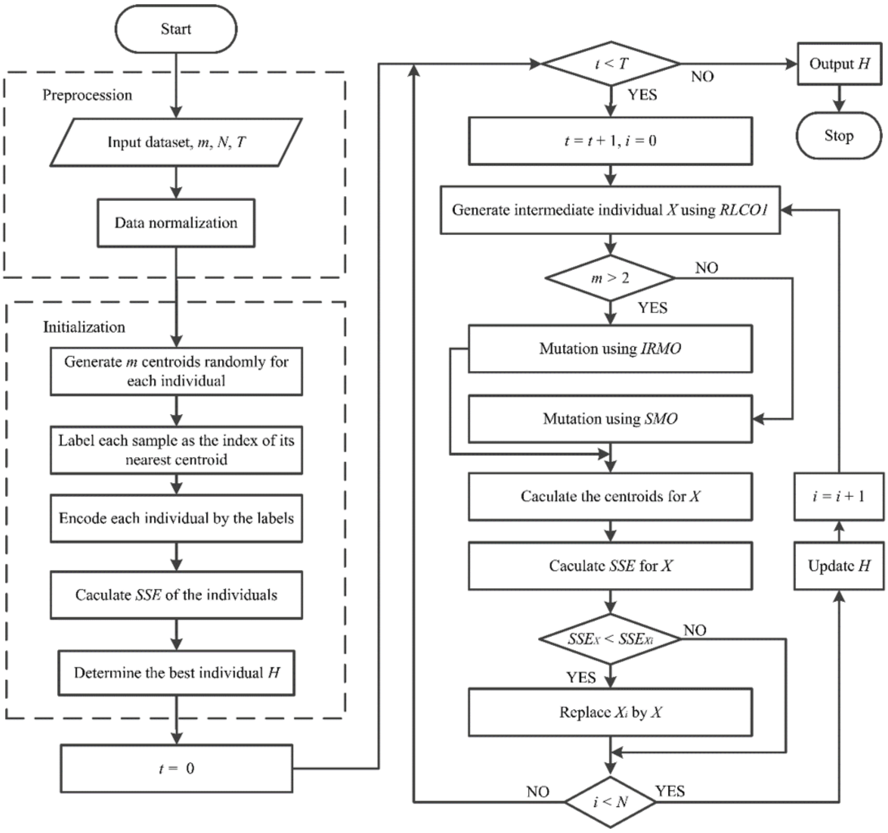

Figure 1. GTOA clustering algorithm process

4 Microseismic signal picking algorithm based on GTOA clustering

4.1 Feature extraction

Feature extraction is a key link in data analysis and machine learning. By selecting and transforming important attributes, it can simplify and enhance algorithm performance while improving computational efficiency. This paper chooses the amplitude, mean, and standard deviation of the signal as its time-domain features. Let $X = \{x_1, x_2, ..., x_n\}$ be the microseismic dataset to be processed. By selecting a continuous set of m points in the collection, the energy of these m points can be calculated as the feature value of the midpoint. The feature functions are shown as 8 , 9 , and 10:

Amplitude feature:

$$A_j = \sum_{i=1+j-m/2}^{j+m/2} X_i^2 \tag{8}$$

https://doi.org/10.5194/egusphere-2025-2884
Preprint. Discussion started: 14 July 2025

Mean feature:

$$\mu_j = \frac{1}{m}\sum_{i=1+j-m/2}^{j+m/2} X_i^2 \quad (9)$$

Standard deviation feature:

$$\sigma_j = \sqrt{\frac{1}{m}\sum_{i=1+j-m/2}^{j+m/2} (X_i - \mu_i)^2} \quad (10)$$

Here, i = 1, 2, . . . , n, j = m/2, . . . , n-m/2. The corresponding feature vector for the sample point X_j in the microseismic data is (A_j, μ_j, σ_j).

4.2 Microseismic signal picking process

The basic idea of the microseismic signal picking algorithm based on GTOA clustering is to regard waveform components as different categories, specifically noise data and useful signal data. After using GTOA to filter out the sample point index range that contains microseismic events, the AIC algorithm is then employed for precise signal picking.

The specific implementation is as follows: after feature extraction, the GTOA clustering algorithm is used to categorize all sample points into two classes: noise (Class 0) and microseismic signals (Class 1). Since microseismic events usually have a short duration and weak signals, the class with a larger number of sample points is determined to be noise, while the class with fewer sample points is deemed as valid signals. Furthermore, the distance from each sample point to the centroids of the two classes is calculated and normalized to the range [0,1] to quantify the tendency of sample points towards each category: the closer the distance to the Class 0 centroid, the more significant the noise characteristics; conversely, the closer the distance to the Class 1 centroid, the more likely it is to be a microseismic signal. Based on this, a threshold of 0.5 is set to filter out sample points that are closer to the Class 1 centroid (i.e., sample points with distances greater than 0.5 are marked as valid microseismic signals), outlining the index range of sample points containing microseismic events. Finally, to accurately locate the arrival time of the direct wave of the microseismic event, the Akaike Information Criterion (AIC) is introduced. This criterion, derived from information theory, identifies the moment of statistical feature changes by balancing the goodness of fit of the model with its complexity. When the order M of the AR model is much smaller than the number of points, the AIC calculation value can be expressed by the following formula 11:

$$A(k) = k log_{10}[\mathrm{var}(x[1,k])] + (N-k-1) log_{10}[\mathrm{var}(x[k+1,N])] + C \quad (11)$$

Here, var represents the variance of the data sequence.

The data within the range is input into the AIC algorithm for picking the arrival time. Due to the significant statistical characteristics of noise signals compared to microseismic signals, the fitting quality of these two types of signals under the least squares criterion is the worst, resulting in a minimal AIC value. Therefore, the moment with the largest variation in AIC value before and after is considered the arrival time of the direct wave of the microseismic signal. The complete processing flow for microseismic signal picking is illustrated in the figure 2 shown.

https://doi.org/10.5194/egusphere-2025-2884
Preprint. Discussion started: 14 July 2025

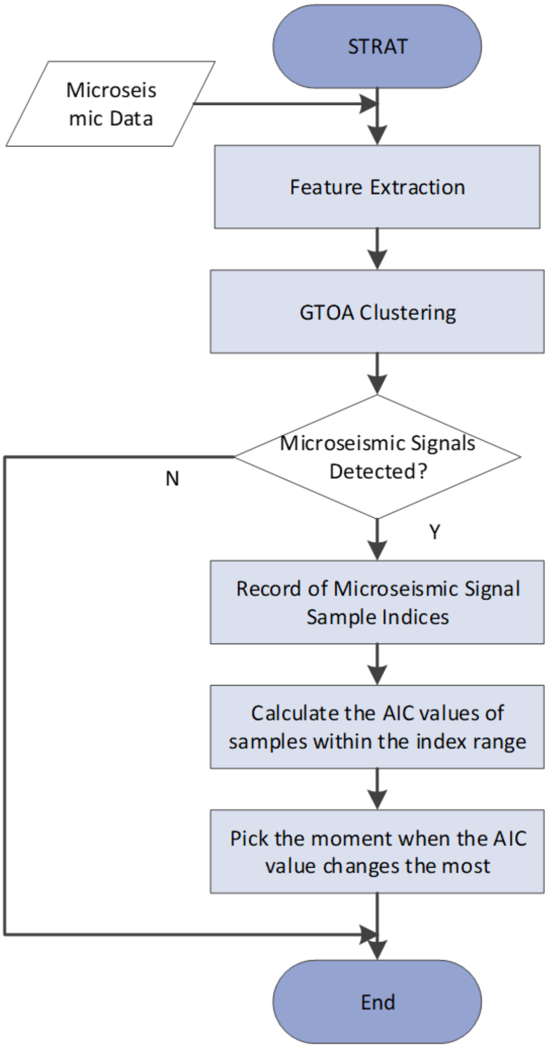

Figure 2. The process of microseismic signal picking algorithm based on GTOA clustering

205 ## 5 GTOA clustering test

5.1 Performance testing of GTOA clustering

The clustering performance of GTOA is tested using four widely cited UCI datasets: Seeds(Charytanowicz and Lukasik, 2010), Glass Identification(German, 1987), Heart Disease(Janosi and Detrano, 1989), and Ionosphere(Sigillito and Baker, 1989), to demonstrate the feasibility and superiority of GTOA clustering. A comparison is made between GTOA clustering and other clustering methods based on evolutionary algorithms, including commonly used GA, PSO, and DE, as well as recent algorithms SCA(Duan and Yu, 2023), WOA(Shu et al., 2024), and TSA(Aslan et al., 2023).

https://doi.org/10.5194/egusphere-2025-2884
Preprint. Discussion started: 14 July 2025

5.1.1 Parameter Settings

To enhance the clustering performance of the GTOA algorithm, this paper employs a dynamic adaptive mutation probability strategy. To optimize the configuration of algorithm parameters, the focus is on examining the impact mechanism of the upper limit of mutation probability P_{max} on clustering performance. Systematic parameter sensitivity experiments were designed to test 8 sets of values forP_{max} within the range of 0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.008, 0.01 on typical datasets, aiming to establish a more reasonable P_{max} parameter value.

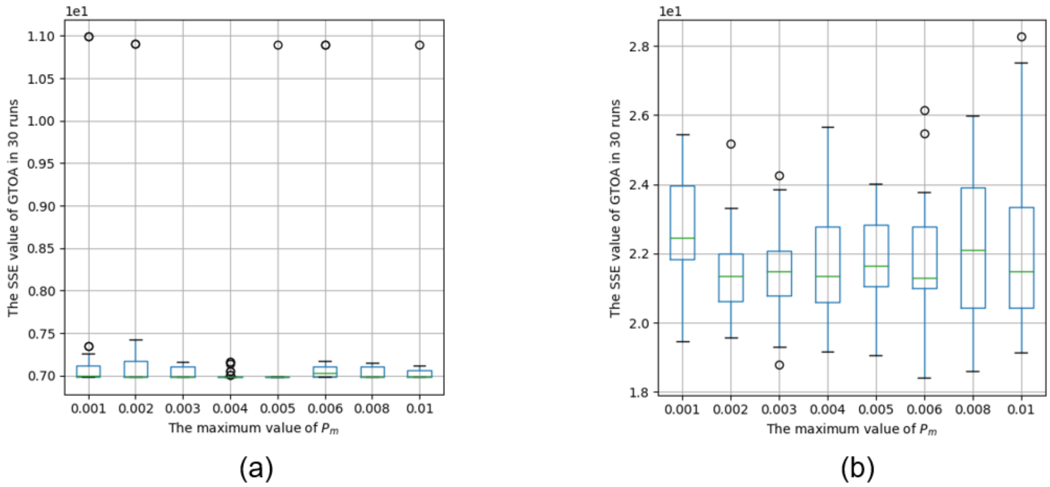

Figure 3. The best SSE obtained in each run for (a) Seeds, (b) Glass Identification dataset

Figure3 displays a box plot of the best SSE obtained by GTOA at different P_{max} values. It can be observed that whenP_{max} is set to 0.003, 0.004, or 0.005, GTOA achieves better computational results. Since the average rank is best at P_{max} = 0.004, this value is selected.

To ensure a fair comparison, the overall size of all algorithms is set to 30, and the maximum number of iterations is set to 500. Each algorithm runs independently 30 times on each benchmark dataset. Detailed explanations of the settings for other parameters are provided in Table 1.

5.1.2 Validation Evaluation Method

Due to the unsupervised nature of clustering, a reasonable validation of clustering results is also a key issue, requiring examination of intra-class similarity and inter-class differences from multiple dimensions. To this end, this paper adopts a comprehensive evaluation strategy: using the sum of squared errors (SSE) as the evaluation metric within the algorithm while additionally introducing external evaluation metrics—Purity and Entropy—to analyze the clustering results. The calculation formulas are as follows. Here, n denotes the total number of samples, $V = \{v_1, v_2, ..., v_K\}$ represents the cluster divisions obtained from clustering, and $S = \{s_1, s_2, ..., s_J\}$ represents the true category divisions.

https://doi.org/10.5194/egusphere-2025-2884
Preprint. Discussion started: 14 July 2025

Table 1. Parameter settings of the algorithms

Algorithm	Parameter	Value
GTOA	P_{max}	0.004
	P_{min}	0.001
GA	Crossover probability	0.8
	Mutation probability	0.02
PSO	C_1	2
	C_2	2
	W_{max}	0.9
	W_{min}	0.2
DE	Scale factor	0.3
	Crossover probability	0.5
SCA	A	2
WOA	A_{max}	2
	A_{min}	0
	B	1
TSA	P_{min}	1
	P_{max}	4

The purity calculation employs the mode labeling method: the most frequently occurring sample point in each cluster is used as the label for that cluster, and the proportion of correctly classified samples to the total number of sample points represents the purity value, P $\in$ (0, 1]. A purity value closer to 1 indicates better clustering results.The formula is shown as 12.

$$P(V,S) = \frac{1}{n}\sum_{k=1}^{K}\max_{j}|v_k \cap s_j| \tag{12}$$

Entropy, as an important indicator for evaluating clustering quality, is primarily used to quantify the level of uncertainty in partition results. The entropy value En $\in$ [0, logK] indicates that a smaller entropy value signifies higher certainty in the clustering results, meaning that similar samples within the same class are more alike, while the differences between samples from different classes are more pronounced.The entropy calculation formulas are shown as 13 and 14.

$$En(V) = \sum_{k=1}^{K}\frac{|v_k|}{n}E(v_k) \tag{13}$$

$$E(v_k) = -\sum_{j=1}^{J}Q(v_k)logQ(v_k) = -\sum_{j=1}^{J}\frac{|v_k \cap s_j|}{|v_k|}log\frac{|v_k \cap s_j|}{|v_k|} \tag{14}$$

https://doi.org/10.5194/egusphere-2025-2884
Preprint. Discussion started: 14 July 2025

In the formula, $Q(v_k)$ represents the probability of an object randomly selected from v_k falling into class s_j, and $E(v_k)$ indicates the entropy of each cluster.

5.1.3 Comparison results with other clustering methods based on evolutionary algorithms

Tables 2, 3, and 4 show the best values (Best), average values (Mean), and standard deviations (Std) of SSE, purity, and entropy obtained by seven evolutionary methods for clustering results. The best results are highlighted in bold.

Table 2. SSE of the clustering results obtained by GTOA and other EAs

Dataset	SSE	GA	PSO	DE	SCA	WOA	TSA	GTOA
Seeds	Best	22.0329	**22.0244**	**22.0244**	26.7390	22.2033	22.4164	**22.0244**
	Mean	22.0485	**22.0251**	22.0727	31.0926	23.7253	24.6987	22.4968
	Std	0.0143	**0.0010**	0.1082	2.6955	2.8148	4.4678	0.9262
Glass Identification	Best	18.9222	25.6663	19.0009	32.8429	30.6933	27.5843	**18.7992**
	Mean	**20.5032**	30.8366	20.9883	37.3151	32.8573	29.9573	21.5568
	Std	1.5157	1.9460	1.3946	2.0842	1.4757	1.2498	**1.2087**
Heart Disease	Best	312.6499	312.6799	312.6200	333.7129	314.9731	315.5890	**312.6200**
	Mean	317.5762	316.3073	313.3627	350.7146	332.8389	325.3657	**312.6200**
	Std	8.1271	5.7538	2.1010	9.8872	15.1863	8.3870	**0.0001**
Ionosphere	Best	629.8421	648.0609	629.1985	795.7066	677.3979	659.7427	**628.8960**
	Mean	630.5990	675.1144	636.2899	845.8288	729.6238	703.1217	**628.8965**
	Std	0.3894	17.6429	5.8786	25.8897	48.4605	53.2936	**0.0019**

As shown in Table 2, for the SEE results of various algorithms, GTOA outperforms other algorithms on the Heart Disease and Ionosphere datasets. For the Seeds dataset, it achieves the best optimal SSE value, and for the Glass Identification dataset, it has the best optimal value with the smallest standard deviation.

As presented in Table 3, GTOA attains both the highest average purity and the lowest standard deviation in clustering results on the Heart Disease and Ionosphere datasets, outperforming the other algorithms. For the Glass Identification dataset, GTOA demonstrates superior performance, with both its optimal and average purity values exceeding those of the competing methods. Additionally, GTOA achieves the highest optimal purity value on the Seeds dataset.

According to the data in Table 4, GTOA stands out with the best average entropy and the least standard deviation on the Heart Disease and Ionosphere datasets. Furthermore, it achieves a superior optimal value compared to other algorithms on the Glass Identification dataset.

From the comparative results above, it can be concluded that the proposed GTOA demonstrates strong competitiveness across the four datasets, with overall performance slightly better than GA and DE, and significantly better than PSO, SCA, WOA, and TSA.

https://doi.org/10.5194/egusphere-2025-2884
Preprint. Discussion started: 14 July 2025

Table 3. Purity of the clustering results obtained by GTOA and other EAs

Dataset	Purity	GA	PSO	DE	SCA	WOA	TSA	GTOA
Seeds	Best	0.8905	0.8905	0.8905	0.9048	0.9143	0.9000	**0.9238**
	Mean	0.8894	0.8887	**0.8897**	0.8451	0.8724	0.8506	0.8819
	Std	0.0020	0.0023	**0.0018**	0.0670	0.0525	0.0842	0.0286
Glass Identification	Best	0.5561	0.5140	0.5374	0.5374	0.5234	0.5187	**0.6355**
	Mean	0.5249	0.4868	0.5181	0.4964	0.4854	0.4949	**0.5405**
	Std	0.0129	0.0174	**0.0076**	0.0176	0.0181	0.0159	0.0259
Heart Disease	Best	0.8000	0.8037	0.8000	**0.8333**	0.8185	0.8148	0.7963
	Mean	0.7468	0.7801	0.7941	0.7599	0.7240	0.7253	**0.7963**
	Std	0.0864	0.0522	0.0151	0.0599	0.0807	0.0832	**0.0000**
Ionosphere	Best	0.7123	0.7208	0.7123	0.7464	**0.8006**	0.7607	0.7123
	Mean	0.7104	0.6961	0.7053	0.6692	0.7080	0.6933	**0.7121**
	Std	0.0014	0.0109	0.0064	0.0295	0.0345	0.0234	**0.0007**

Table 4. Entropy of the clustering results obtained by GTOA and other EAs

Dataset	Entropy	GA	PSO	DE	SCA	WOA	TSA	GTOA
Seeds	Best	0.3588	0.3588	0.3588	**0.2869**	0.3155	0.3346	0.2887
	Mean	0.3611	0.3624	**0.3604**	0.3956	0.3690	0.4024	0.3685
	Std	0.0042	0.0048	**0.0037**	0.0815	0.0606	0.0910	0.0302
Glass Identification	Best	0.9621	1.0218	0.9560	0.9854	1.0328	0.9721	**0.9427**
	Mean	**1.0097**	1.1162	1.0123	1.0836	1.1227	1.0775	1.0163
	Std	**0.0265**	0.0444	0.0363	0.0438	0.0492	0.0444	0.0386
Heart Disease	Best	0.4942	0.4896	0.4942	**0.4461**	0.4730	0.4778	0.4999
	Mean	0.5392	0.5140	0.5021	0.5317	0.5631	0.5582	**0.4999**
	Std	0.0715	0.0446	0.0181	0.0503	0.0670	0.0660	**0.0000**
Ionosphere	Best	0.5623	0.5622	0.5623	0.5626	**0.4888**	0.5487	0.5623
	Mean	0.5642	0.5835	0.5703	0.6152	0.5725	0.5855	**0.5625**
	Std	0.0016	0.0105	0.0072	0.0213	0.0295	0.0233	**0.0008**

Figure 4 illustrates the average convergence curves of various algorithms, allowing for the observation of the convergence process of all algorithms, which displays the average of the optimal SSE values throughout the iterations. From the convergence graph, it can be seen that GTOA demonstrates stable convergence capability, being able to find better solutions as the number of iterations increases. The convergence speed of GTOA is noticeably faster than that of PSO and SCA, and in some datasets, GTOA's convergence speed also surpasses that of GA and DE.

https://doi.org/10.5194/egusphere-2025-2884
Preprint. Discussion started: 14 July 2025

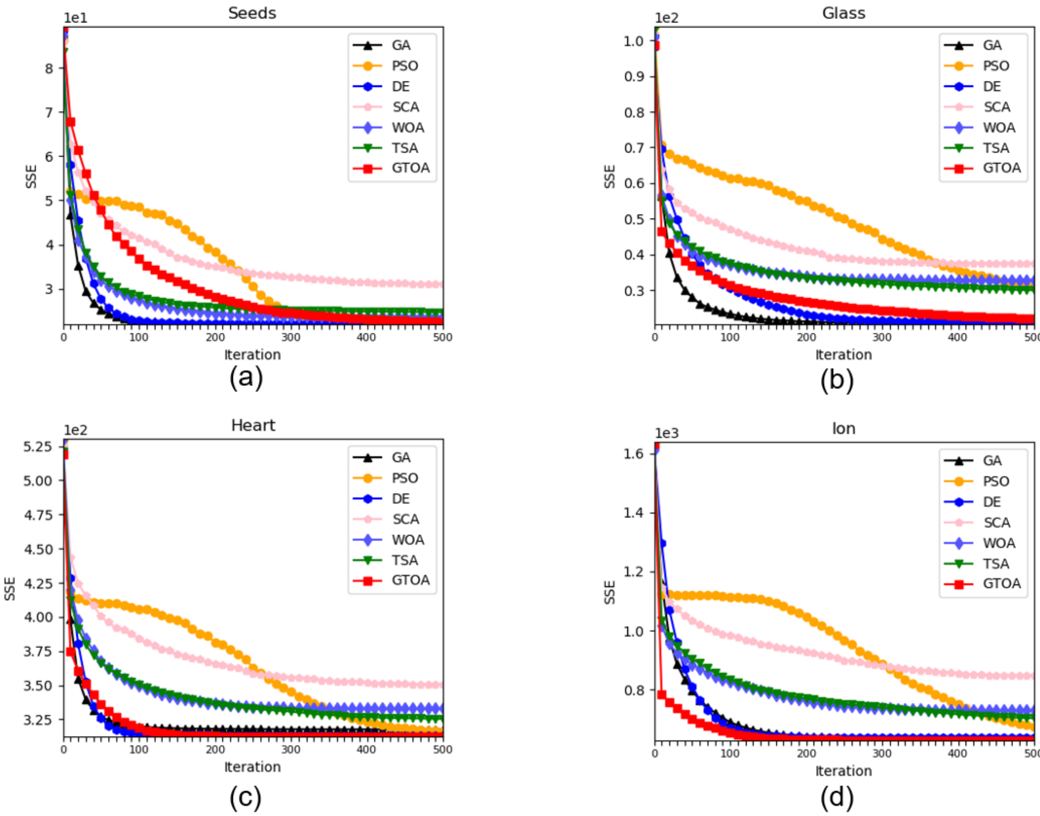

Figure 4. Mean convergence curves of the algorithms for (a) Seeds (b) Glass Identification (c) Heart Disease (d) Ionosphere dataset

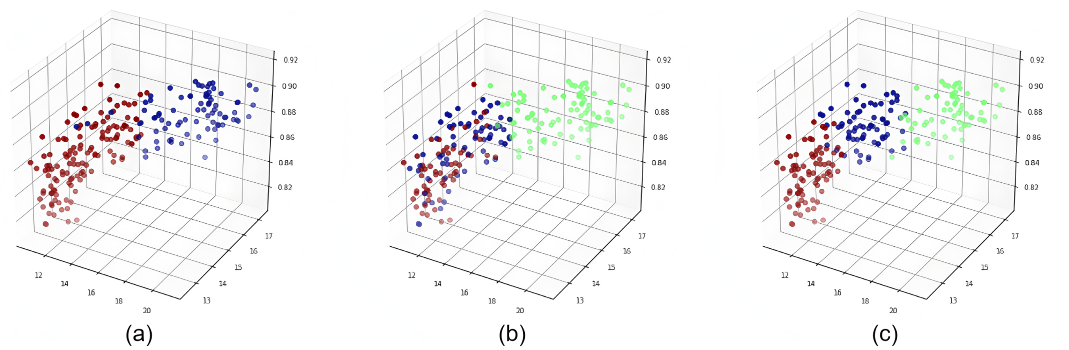

Figure 5. GTOA clustering results with different iteration counts (a) 0, (b) 300, (c) 500

5.2 Demonstration of the dynamic optimization effects of GTOA clustering

The three-dimensional scatter plot in Figure 5 dynamically demonstrates the GTOA clustering process. In the figure, different colors represent different clusters, and the axes correspond to the three feature dimensions of the data points. It can be observed

https://doi.org/10.5194/egusphere-2025-2884
Preprint. Discussion started: 14 July 2025

that the clustering results obtained from random initialization exhibit significant differences from the true classes. However, as the number of iterations increases, GTOA can gradually assign similar data points to the same category, ultimately achieving clustering results that are highly consistent with the true labels of the dataset.

6 Effect testing of the microseismic signal picking algorithm based on GTOA clustering

In order to verify the effectiveness of the proposed microseismic signal picking algorithm, test experiments were conducted. In this experiment, a Ricker wavelet with a frequency of 200 Hz was used as the source signal, and a time sampling interval of 1 ms was set for the numerical simulation of the seismic wave field. The Ricker wavelet is a zero-phase wavelet that is widely used in seismic wave field simulation.(Lauro et al., 2022) To simulate the interference that microseismic signals may experience in a real environment, different levels of noise were added to the generated microseismic waveform data to test the proposed algorithm's performance in picking microseismic signals under varying noise levels. For the algorithm put forward in this paper, the population size of the GTOA clustering was set to 30, with a maximum iteration count of 200.

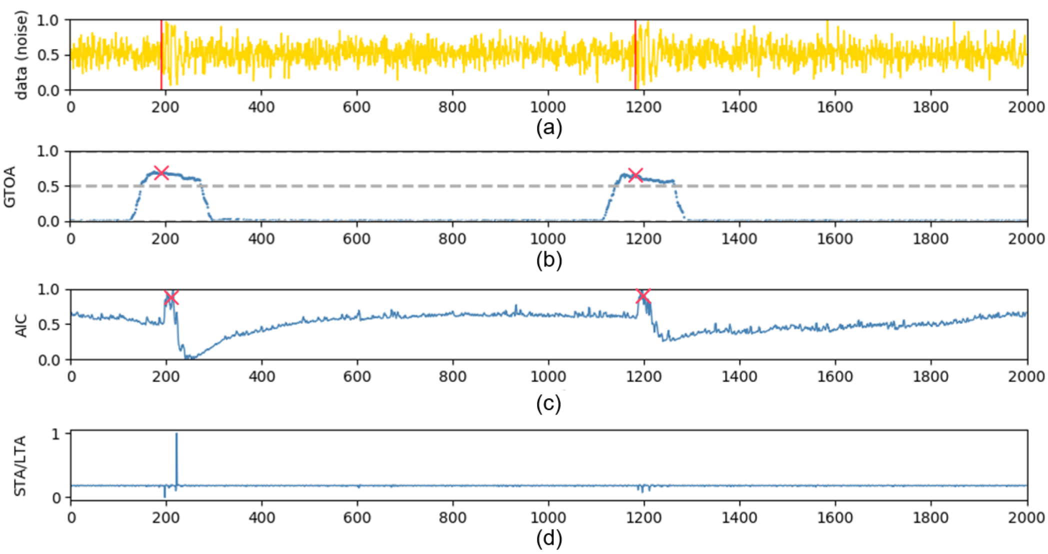

Figure 6. Signal picking results at SNR=-7 dB (a) Microseismic data at SNR = -7 dB (b) pickup results based on GTOA clustering algorithm (c) traditional AIC pickup results (d) traditional STA/LTA pickup results

Figure 6 presents a comparison of the microseismic signal picking effect at a signal-to-noise ratio of -7dB. The simulated noisy microseismic data contains two microseismic signals, indicated by red lines in the figure. Despite the high noise level, the GTOA clustering algorithm designed in this paper is still able to accurately identify the effective signal range, and when combined with the AIC algorithm, it can effectively perform signal picking. In contrast, the traditional STA/LTA method may result in missed detections, only being able to detect the first microseismic signal, while the conventional AIC algorithm experiences delays in picking time.

https://doi.org/10.5194/egusphere-2025-2884
Preprint. Discussion started: 14 July 2025

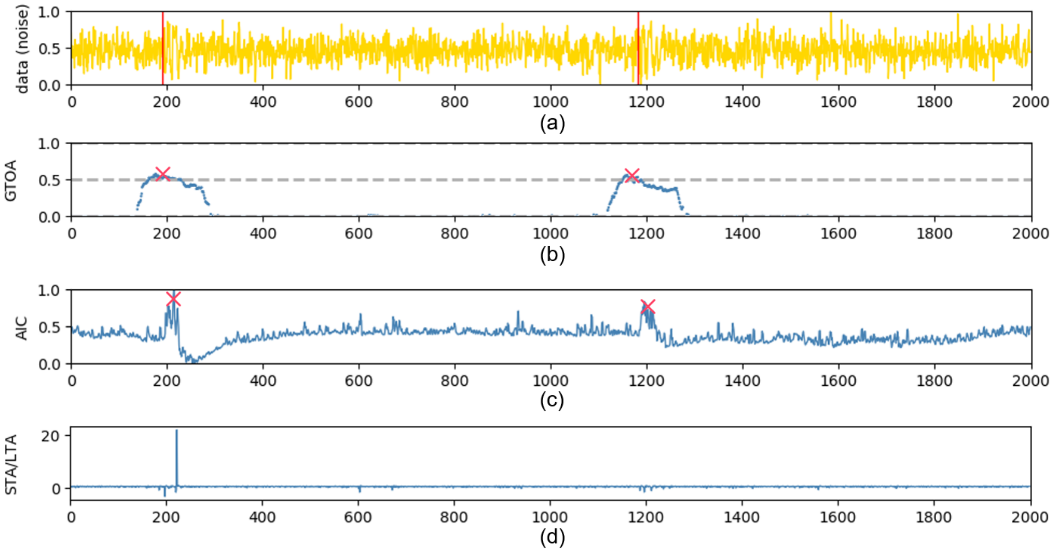

Figure 7. Signal picking results at SNR=-10 dB (a) Microseismic data at SNR = -10 dB (b) pickup results based on GTOA clustering algorithm (c) traditional AIC pickup results (d) traditional STA/LTA pickup results

Figure 7 further illustrates the comparison of microseismic signal picking effects at a signal-to-noise ratio of approximately -10dB. Despite the further reduction in the signal-to-noise ratio, which increases the difficulty of identification, the algorithm designed in this paper can still identify the effective signal range and perform relatively accurate signal picking.Compared to the traditional STA/LTA or AIC algorithms, there is a significant improvement in accuracy.

It can be seen that the proposed method is able to accurately pick up the correct microseismic signals, and the precision of signal picking is high. It can adapt to signal inputs with lower signal-to-noise ratios, making it a reliable signal picking method.

7 Conclusions

This paper designs a microseismic signal picking algorithm based on GTOA clustering, which innovatively uses the GTOA clustering algorithm as its foundation, combined with the Akaike Information Criterion (AIC) for microseismic signal picking. The GTOA clustering algorithm is the core of the proposed algorithm. For the clustering problem, this paper defines a representation method suitable for GTOA solutions, where individual encoding directly represents the partitioning results of the dataset. In each iteration, the centroids' coordinates are re-obtained using the concept of means, effectively reducing the overall computational complexity compared to other evolutionary algorithms that optimize centroid coordinates.Building on this, the search strategy was adjusted, utilizing the best individuals to guide the evolutionary process and accelerate convergence; the mutation probability was dynamically adjusted to balance exploration and exploitation, thereby improving solution quality.

Experimental results on four UCI datasets indicate that GTOA clustering is an effective clustering algorithm. Compared to other evolutionary algorithm-based clustering methods, this approach is more efficient and yields better solutions. Furthermore,

https://doi.org/10.5194/egusphere-2025-2884
Preprint. Discussion started: 14 July 2025

under different signal-to-noise ratio conditions, comparative experiments between the microseismic signal picking algorithm based on GTOA clustering and traditional picking methods demonstrate that the proposed method can effectively pick microseismic signals and achieve greater precision in picking valid microseismic signals.

Data Availability

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

Author Contribution

This research was designed, tested, and implemented by the authors of the paper. The full text was designed and implemented by HZ. ZL and LC conducted testing and debugging.QZ and YW carried out revision and correction during the completion of the article.

Competing Interests

The authors declare that they have no conflict of interest.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFC2801404 and Grant 2022YFF0706202, in part by the National Natural Science Foundation of China under Grant 42074155, in part by the Fundamental Research Funds for the Central Universities under Grant 2-9-2023-011, and in part by the Key Research Program of the Chinese Academy of Sciences under Grant KGFZD-145-22-06-02.

References

Aslan, S., Kızıloluk, S., and Sert, E.: TSA-CNN-AOA: Twitter sentiment analysis using CNN optimized via arithmetic optimization algorithm, Neural Computing and Applications, 35, 10 311–10 328, 2023.

Camacho-Villalón, C. L., Dorigo, M., and Stützle, T.: PSO-X: A component-based framework for the automatic design of particle swarm optimization algorithms, IEEE Transactions on Evolutionary Computation, 26, 402–416, 2021.

Charytanowicz, Magorzata, N. J. K. P. K. P. and Lukasik, S.: Seeds, UCI Machine Learning Repository, 2010.

CHEN, B.-r., FENG, X.-t., FU, Q.-q., WANG, B., ZHU, X.-h., LI, T., LU, C.-p., and XIA, H.: Integration and high precision intelligence microseismic monitoring technology and its application in deep rock engineering, Rock and Soil Mechanics, 41, 8, 2020.

Chen, Z., Liu, H., Zhao, J., and Bi, T.: Real-time event detection based on STA/LTA method using field synchrophasor measurements, IEEE Transactions on Power Delivery, 38, 4070–4080, 2023.

Deng, W., Shang, S., Cai, X., Zhao, H., Song, Y., and Xu, J.: An improved differential evolution algorithm and its application in optimization problem, Soft Computing, 25, 5277–5298, 2021.

Duan, Y. and Yu, X.: A collaboration-based hybrid GWO-SCA optimizer for engineering optimization problems, Expert Systems with Applications, 213, 119 017, 2023.

Feng, Q., Han, L., Zhao, B., and Li, Q.: Microseismic events recognition via joint deep clustering with residual shrinkage dense network, IEEE Transactions on Geoscience and Remote Sensing, 61, 1–11, 2023.

Gad, A. G.: Particle swarm optimization algorithm and its applications: a systematic review, Archives of computational methods in engineering, 29, 2531–2561, 2022.

German, B.: Glass Identification, UCI Machine Learning Repository, 1987.

He, Y. and Wang, X.: Group theory-based optimization algorithm for solving knapsack problems, Knowledge-Based Systems, 219, 104 445, 2021.

Ileberi, E., Sun, Y., and Wang, Z.: A machine learning based credit card fraud detection using the GA algorithm for feature selection, Journal of Big Data, 9, 24, 2022.

Janosi, Andras, S. W. P. M. and Detrano, R.: Heart Disease, UCI Machine Learning Repository, 1989.

Jiang, R., Dai, F., Liu, Y., and Li, A.: A novel method for automatic identification of rock fracture signals in microseismic monitoring, Measurement, 175, 109 129, 2021.

Jin, H., Li, L., Cheng, S., and Li, X.: Time-arrival pickup method of tunnel water inrush microseismic signals based on kurtosis value and AIC method, Tunnelling and Underground Space Technology, 154, 106 135, 2024.

Lauro, S. E., Baniamerian, J., Cosciotti, B., Mattei, E., and Pettinelli, E.: Loss tangent estimation from ground-penetrating radar data using Ricker wavelet centroid-frequency shift analysis, Geophysics, 87, H1–H12, 2022.

Li, H., Yang, Z., and Yan, W.: An improved AIC onset-time picking method based on regression convolutional neural network, Mechanical Systems and Signal Processing, 171, 108 867, 2022a.

Li, Z., Zhang, Q., and He, Y.: Modified group theory-based optimization algorithms for numerical optimization, Applied Intelligence, 52, 11 300–11 323, 2022b.

Liu, J.-p., Si, Y.-t., Wei, D.-c., Shi, H.-x., and Wang, R.: Developments and prospects of microseismic monitoring technology in underground metal mines in China, Journal of Central South University, 28, 3074–3098, 2021.

https://doi.org/10.5194/egusphere-2025-2884
Preprint. Discussion started: 14 July 2025

Ma, H., Wei, H., Tian, Y., Cheng, R., and Zhang, X.: A multi-stage evolutionary algorithm for multi-objective optimization with complex constraints, Information Sciences, 560, 68–91, 2021.

Mirjalili, S. and Mirjalili, S.: Genetic algorithm, Evolutionary algorithms and neural networks: Theory and applications, pp. 43–55, 2019.

QIU, L. and LI, C.: STA/LTA method for picking up the first arrival of natural seismic waves and its improvement analysis, Progress in Geophysics, 38, 1497–1506, 2023.

Shu, T., Pan, Z., Ding, Z., and Zu, Z.: Resource scheduling optimization for industrial operating system using deep reinforcement learning and WOA algorithm, Expert Systems with Applications, 255, 124 765, 2024.

Sigillito, V., W. S. H. L. and Baker, K.: Ionosphere, UCI Machine Learning Repository, 1989.

Song, Y., Wu, Y., Guo, Y., Yan, R., Suganthan, P. N., Zhang, Y., Pedrycz, W., Das, S., Mallipeddi, R., Ajani, O. S., et al.: Reinforcement learning-assisted evolutionary algorithm: A survey and research opportunities, Swarm and Evolutionary Computation, 86, 101 517, 2024.

Wang, R., Zhang, Z., Ng, W. W., and Wu, W.: An improved group theory-based optimization algorithm for discounted 0-1 knapsack problem, Advances in Computational Intelligence, 1, 1–11, 2021a.

Wang, S., Shi, Q., Wang, S., Shen, Y., Sun, Q., and Cai, Y.: Resource property and exploitation concepts with green and low-carbon of tar-rich coal as coal-based oil and gas, Journal of China Coal Society, 46, 1365–1377, 2021b.