Rock and Soil Mechanics

Volume 43 | Issue 6 Article 6

8-1-2022

Dominant partitioning method of rock mass discontinuity based on DBSCAN selective clustering ensemble

Hua-jin ZHANG
Faculty of Land Resources Engineering, Kunming University of Science and Technology, Kunming, Yunnan 650093, China

Shun-chuan WU
School of Civil and Resources Engineering, University of Science and Technology Beijing, Beijing 100083, China, huajinzhang1001@163.com

Long-qiang HAN
Faculty of Land Resources Engineering, Kunming University of Science and Technology, Kunming, Yunnan 650093, China

Rock and Soil Mechanics 2022 43(6): 1585−1595 ISSN 1000−7598
https: //doi.org/10.16285/j.rsm.2021.6582 rocksoilmech.researchcommons.org/journal

Dominant partitioning method of rock mass discontinuity based on DBSCAN selective clustering ensemble

ZHANG Hua-jin[1], WU Shun-chuan[1, 2], HAN Long-qiang[1]

1. Faculty of Land Resources Engineering, Kunming University of Science and Technology, Kunming, Yunnan 650093, China
2. School of Civil and Resources Engineering, University of Science and Technology Beijing, Beijing 100083, China

Abstract: For the problems existing in the traditional single discontinuity (structural plane) based clustering model, such as the risk of misclassification or omission and the inability to identify noise and isolated values, a dominant partitioning method of rock mass discontinuity based on selective clustering ensemble using density-based spatial clustering of applications with noise (DBSCAN) algorithm is proposed. Firstly, the spatial coordinate transformation is performed with the attitude of discontinuity, and the sine of the angle between the unit normal vectors is defined as similarity measurement. Then, a certain number of different base clusters are constructed based on the DBSCAN algorithm, with the selective clustering ensemble technology, some excellent base clusters are selected. Finally, the consistent ensemble technology is used to fuse these base clusters to generate a highly reliable selective clustering ensemble result. The DIPS software data set and the discontinuity survey result in the dam site area of Songta hydropower station are used to test the feasibility and effectiveness of the proposed method. The research results show that the clustering effect of the proposed method is significantly better than that of common clustering algorithms. The clustering results are objective and reasonable. It not only effectively identifies noise and isolated values, but also overcomes the shortcomings of over-segmentation or under-segmentation of the single discontinuity based clustering model. The research results are valuable in accurately determining the dominant group of discontinuity.
Keywords: rock mass discontinuity; dominant attitude; clustering ensemble; density-based spatial clustering of applications with noise (DBSCAN); silhouette coefficient

1 Introduction

As an important component of the rock mass, the structural plane (discontinuity) greatly controls the mechanical properties and engineering stability of the rock mass[1]. The complexity of the formation of structural planes determines that their distribution has not only certain regularity, but also randomness and uncertainty. The rock mass structural planes in the natural state are usually distributed in groups. The partitioning of random structural plane attitude is a basis for characterizing the distribution law of structural planes and it is also a prerequisite for the simulation of randomly jointed rock mass and the rock mass stability evaluation. The analysis results are of considerable significance for the strength parameter selection, mechanical property study, and engineering stability evaluation of rock mass[2−3].

At present, the conventional analysis methods for structural plane attitude such as rose diagram, pole diagram and isodensity diagram are widely used in the projects. These methods are simple and intuitive, but the partitioning results of dominant attitude depend on the professional knowledge and experience of analysts, which is subjective to a certain extent, and it is difficult to obtain accurate results in the areas with highly developed structural planes[4−5]. In order to solve the shortcomings of conventional statistical analysis methods of structural plane, numerous scholars tried to objectively analyze the structural planes from the sample data itself with the help of the clustering analysis method. Since Shanley et al.[6] first proposed the clustering analysis method for the structural plane attitude, great progress has been made in the research on the clustering of structural planes. *K*-means is the most widely used algorithm in the dominant partitioning of structural planes. On this basis, Hammah et al.[7] introduced the membership function and proposed the fuzzy C-means (FCM) algorithm to partition the dominant attitude of structural planes. Cui et al.[8] and Li et al.[9] used heuristic algorithms such as genetic algorithm and ant colony algorithm to continuously optimize and search the global optimal solution, and finally solved the problems that FCM clustering results are sensitive to the initial clustering center value and easy to fall into the local optimal solution. However, *K*-means and its derived FCM clustering algorithm need to determine the initial clustering center in advance. In order to overcome this problem, Zhang et al.[5] proposed a hierarchical clustering analysis method for partitioning of structural plane attitude. Jimenez-Rodriguez et al.[10] used the spectrum-based clustering algorithm to partition the structural plane attitude without determining the clustering center in advance, and the algorithm has fast convergence speed and good clustering effect. However, these algorithms still need to determine the number of

Received: 17 September 2021 Revised: 3 March 2022
This work was supported by the National Natural Science Foundation of China (51934003) and the Program of Yunnan Innovation Team (202105AE160023).
First author: ZHANG Hua-jin, male, born in 1996, PhD, mainly engaged in the research on rock mechanics and rock mass stability analysis.
E-mail: huajinzhang1001@163.com
Corresponding author: WU Shun-chuan, male, born in 1969, PhD, Professor, mainly engaged in the teaching and research on mining engineering and geotechnical engineering.

structural plane groups manually in advance, and it is difficult to identify the noise and isolated values in the structural plane attitude. Density-based spatial clustering of applications with noise (DBSCAN)[11] algorithm obtains clustering results through the connectivity between samples from the perspective of sample density. It is capable of filtering noise and isolated values. In recent years, it has been applied in the field of structural plane classification of rock mass point cloud data[12], but its application to the dominant attitude partitioning of structural plane is seldom reported.

On the other hand, each model has its own optimization standard and assumption. Because the structural plane attitude is of great randomness, the assumption may not conform to the real distribution of structural planes. Therefore, a single model may not be able to obtain an accurate and effective result of structural plane attitude clustering analysis. Especially for the structural plane noise and the boundary points between groups, there is a great risk of false selection and mis-selection. Integrated learning[13] reveals the essential characteristics of data sets from different levels by combining multiple base clusters, which is an effective approach to solving the poor clustering effect of a single model. The performance of the clustering ensemble model depends on the quality of the base cluster, which is often difficult to ensure. Therefore, Fern et al.[14] proposed a concept of selective clustering ensemble, which ensures the final clustering effect by integrating some base clusters with good quality and large difference from each other. The selective clustering ensemble can usually achieve a better clustering effect than a single model, and has been widely used in many fields, but it has not been reported in the partitioning of dominant structural plane attitude.

Therefore, in the light of the shortcomings of the current dominant attitude partitioning of structural plane, a method for dominant attitude partitioning of structural plane using the DBSCAN-based selective clustering ensemble algorithm is proposed. Based on the spatial coordinate transformation of structural plane attitude, a certain number of different base clusters are constructed with the DBSCAN algorithm. With the aid of selective clustering ensemble technology, some excellent base clusters are selected. Then, these base clusters are integrated and complemented with the consistent ensemble technology, and a highly reliable selective clustering ensemble result is obtained. This method combines the advantages of the DBSCAN algorithm and the selective clustering ensemble algorithm. It can effectively eliminate the noise and isolated values in the attitude data of structural planes, and avoid the misjudgment or mis-selection of a single model. The method is applied to processing the DISP software data set and structural plane survey results in the dam site area of Songta hydropower, and the satisfactory results are obtained, which verifies the feasibility and effectiveness of the proposed method.

2 Similarity measurement of spatial structural plane

During collecting the in-situ structural plane data, the dip direction ($0^\circ \leqslant \alpha \leqslant 360^\circ$) and dip angle ($0^\circ \leqslant \beta \leqslant 90^\circ$) are usually used to represent the attitude of structural plane. The dip direction represents the inclination of the structural plane in space, and the dip angle represents the angle between the structural plane and the horizontal plane. When conducting mathematical analysis on the attitude data of structural planes, it is generally assumed that the rock mass structural plane is three-dimensional, and its attitude is represented by the corresponding unit normal vector $\boldsymbol{n}_i = (x_i, y_i, z_i)$:

$$\left.\begin{aligned} x_i &= \cos\alpha \sin\beta \\ y_i &= \sin\alpha \sin\beta \\ z_i &= \cos\beta \end{aligned}\right\} \tag{1}$$

The dominant partitioning of structural plane attitudes means that the structural planes with similar spatial directions are classified as the same class, while those with distant directions are classified as different classes. Therefore, a suitable distance function must be used to quantitatively describe the similarity between the two structural planes. The shorter the distance is, the better the similarity is. The selected distance function[9] must satisfy symmetry, non-negativity and self-equivalence, i.e. the distance function $D(\boldsymbol{n}_i, \boldsymbol{n}_j)$ of two structural planes $\boldsymbol{n}_i$ and $\boldsymbol{n}_j$ must satisfy $D(\boldsymbol{n}_i, \boldsymbol{n}_j) = D(\boldsymbol{n}_j, \boldsymbol{n}_i)$, $D(\boldsymbol{n}_i, \boldsymbol{n}_j) \geqslant 0$, $D(\boldsymbol{n}_i, \boldsymbol{n}_i) = 0$. At present, there are many distance criteria for measuring the attitude data of structural planes. For example, Cai et al.[2], Li et al.[9] and Song et al.[15] respectively employed the Euclidean distance, spherical distance, and the sine of the angle between unit normal vectors of structural planes as the distance function. Considering the special case that the Euclidean distance and spherical distance cannot identify the similarity of steep structural planes with 180° difference in dip direction, this paper adopts the sine of the angle between unit normal vectors of structural planes[5, 15] as the similarity measurement index of structural plane attitude, as shown in Fig.1, which not only meets the requirements of the above distance function, but also has a clear geometric meaning. The distance $D(\boldsymbol{n}_i, \boldsymbol{n}_j)$ between the two structural planes

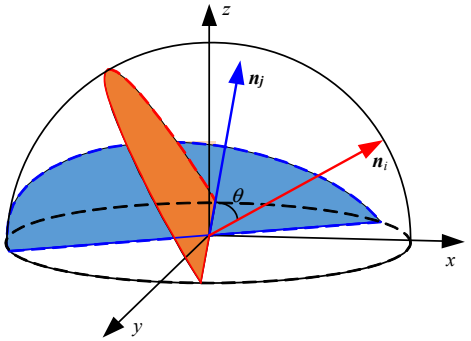


Fig. 1 Schematic diagram of the angle between two unit normal vectors of structural planes

$\boldsymbol{n}_i(x_i, y_i, z_i)$ and $\boldsymbol{n}_j(x_j, y_j, z_j)$ can be expressed as

$$D(\boldsymbol{n}_i, \boldsymbol{n}_j) = \sin\theta = \sqrt{1-(\boldsymbol{n}_i, \boldsymbol{n}_j^{\mathrm{T}})^2} \quad (2)$$

3 Relevant basic theories

3.1 DBSCA algorithm

DBSCAN algorithm[11] is the most famous and representative spatial density clustering algorithm. The algorithm includes two important hyperparameters: radius r and density threshold T_{d}. First, the region with sufficient density is divided into a cluster. If the density of a sample in the r range exceeds T_{d}, the sample is regarded as a core point. The samples around the core point within the radius of r belong to the same cluster as the core point. Samples that cannot be connected within the distance of r of any core point are regarded as noise and do not belong to any cluster, as shown in Fig.2.

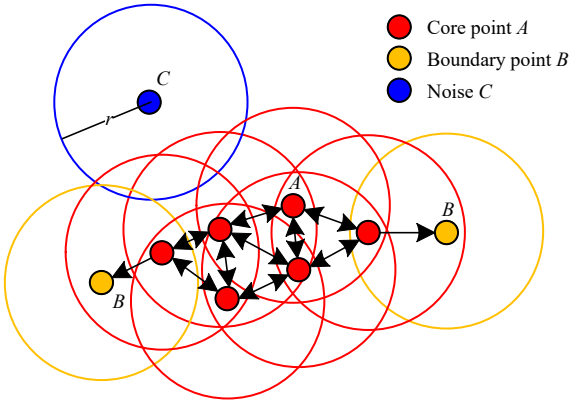


Fig. 2 Illustration of the DBSCAN clustering

DBSCAN algorithm does not need to specify the number of clusters in advance. It is suitable for arbitrary data structures and can identify the noise or isolated values[11]. However, the clustering effect of the DBSCAN algorithm largely depends on the two hyperparameters r and T_{d}. Subtle differences may lead to significantly different clustering results. These two hyperparameters are difficult to specify and mainly depend on experience.

3.2 Clustering ensemble technology

In many cases, the single clustering model is prone to generate over-segmentation or under-segmentation results. Therefore, in order to further improve the clustering performance, Strehl et al.[16] proposed the clustering ensemble technology to solve the same problem by combining multiple base clusters, and used the complementary information between the base clusters to train a highly reliable recognition system to obtain clustering results that are more accurate and robust than a single clustering model. According to the error proof of integrated learning theory by Hansen et al.[17], assuming that there are H mutually independent base clusters, the result error of each base cluster is p. When the voting method is used for ensemble, the error of the ensemble model is

$$E = \sum_{k>H/2}^{H} C_k^H p^k (1-p)^{H-k} \quad (3)$$

It can be seen that when p<0.5, E decreases monotonically with the increase of H. As long as the accuracy of each base cluster is greater than 0.5, E will decrease monotonically with the increase of the number of base clusters participating in integration, and eventually tend to be 0. Meanwhile, the clustering effect of the ensemble model will continue to improve. A key assumption in the above derivation is that the base clusters are independent of each other. However, in fact, these base clusters are trained based on the same task, and thus they are impossible to be independent of each other. Therefore, in order to obtain an ensemble model with a superior clustering effect, the integrated base clusters should be "good but different" as far as possible, i.e. constructing high-quality and different base clusters is necessary for a successful clustering ensemble[18].

3.3 Selective clustering ensemble technology

The quality of the base cluster varies. If some base clusters with poor quality participate in the integration, the clustering effect of the integrated model will be deteriorated. The selective clustering ensemble technology adds a selection stage between the construction and integration of the base cluster, and selects some base clusters with good quality and large difference for integration, which effectively eliminates the interference of the poor base cluster, and further improves the clustering effect[14, 19].

Suppose there are N structural planes in the data set $X=\{x_1, x_2, \cdots, x_N\}$, where $x_n=\{x_{n1}, x_{n2}\}(n=1, 2, \cdots, N)$, in which x_{n1} and x_{n2} represent the dip direction and dip angle, respectively. Perform H times of different cluster analyses on data set X to generate H base cluster results $P=\{P_1, P_2, \cdots, P_H\}$, where $P_i=\{C_i^1, C_i^2, \cdots, C_i^{k_i}\}$ $(i=1, 2, \cdots, H)$ is the cluster result of the ith base cluster, in which k_i represents the number of clusters in the ith base cluster result. Selective clustering ensemble is performed to select partial base clustering results $P'=\{P_1', P_2', \cdots, P_{H'}'\}(1\leqslant H'\leqslant H)$ from P for integration to obtain better clustering effect.

Due to the lack of a priori category information in cluster analysis, the same cluster results may have different class labels. For example, the cluster results $P_1=[1, 1, 2, 2, 3]$ and $P_2=[3, 3, 1, 1, 2]$, the two cluster results have different labels, but represent the same cluster result. Therefore, to solve the problem of inconsistent labels, the base clustering result P_i is transformed into the corresponding incidence matrix $\boldsymbol{M}_i$:

$$\boldsymbol{M}_i(j,k)=\begin{cases}1 & \text{if } P_i^j=P_i^k(1\leqslant j\leqslant N, 1\leqslant k\leqslant N)\\ 0 & \text{otherwise}\end{cases} \quad (4)$$

According to Eq.(4), if a pair of samples j and k appear in the same category, $\boldsymbol{M}_i(j, k)=1$, otherwise $\boldsymbol{M}_i(j, k)=0$, the H base clustering results $P=\{P_1, P_2, \cdots, P_H\}$ are transformed into the corresponding incidence matrix,

$M = \{\boldsymbol{M}_1, \boldsymbol{M}_2, \cdots, \boldsymbol{M}_H\}$. After this transformation, the clustering results of P_1 and P_2 can be consistent, for example, $\boldsymbol{M}_1 = \boldsymbol{M}_2 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$.

If P^* is the true distribution of data set X, for any base clustering result P_i, the accuracy of the ensemble model is defined as

$$N^{(\mathrm{a})} = \frac{\sum_{i=1}^{H} \left\| \boldsymbol{M}_i, \boldsymbol{M}^* \right\|}{H} \tag{5}$$

where $\boldsymbol{M}_i$ and $\boldsymbol{M}^*$ are the incidence matrices of P_i and P^*, respectively; and $\left\| \boldsymbol{M}_i, \boldsymbol{M}^* \right\|$ denotes the similarity of the two matrices. Obviously, the larger the value of $N^{(\mathrm{a})}$ is, the more similar each base clustering result is to the true distribution, and the better the clustering effect will be.

The difference degree of the base cluster is also a essential factor. The average difference degree between the paired base clusters is used as the difference degree of the ensemble model:

$$N^{(\mathrm{d})} = 1 - \frac{\sum_{i=1}^{H} \sum_{j=1,\, j \neq i}^{H} \left\| \boldsymbol{M}_i, \boldsymbol{M}_j \right\|}{H(H-1)} \tag{6}$$

Taking the base clustering result $\bar{P}$ from $S = L^{(\mathrm{a})} \cup L^{(\mathrm{d})}$, $\bar{P} \in P$, $S \subset P$, $S \neq \varnothing$, where $L^{(\mathrm{a})}$ represents the base cluster set with accuracy lower than $N^{(\mathrm{a})}$, and $L^{(\mathrm{d})}$ represents the base cluster set with difference degree lower than $N^{(\mathrm{d})}$. Let $Q = P - \{\bar{P}\}$, indicating that the base clustering result with poor performance $\bar{P}$ is removed from the set P, then the accuracy of base clustering subset Q is

$$N_Q^{(\mathrm{a})} = \frac{\sum_{i=1}^{H} \left\| \boldsymbol{M}_i, \boldsymbol{M}^* \right\| - \left\| \bar{\boldsymbol{M}}, \boldsymbol{M}^* \right\|}{H-1} \tag{7}$$

where $\bar{\boldsymbol{M}}$ is the incidence matrix of $\bar{P}$.

According to Eq.(5), since $\bar{P} \in S$, then $\left\| \bar{\boldsymbol{M}}, \boldsymbol{M}^* \right\| < N^{(\mathrm{a})}$, therefore, it can be obtained that

$$\frac{\sum_{i=1}^{H} \left\| \boldsymbol{M}_i, \boldsymbol{M}^* \right\| - \left\| \bar{\boldsymbol{M}}, \boldsymbol{M}^* \right\|}{H-1} > \frac{H^* N^{(\mathrm{a})} - N^{(\mathrm{a})}}{H-1} = N^{(\mathrm{a})} \tag{8}$$

Similarly, the difference degree $N_Q^{(\mathrm{d})}$ of Q is greater than that of P. From the above proof results, it can be seen that if the base clustering results with poor accuracy and difference degree $\bar{P}$ ($\bar{P} \in S$) are removed from the base clustering result set P, the accuracy $N_Q^{(\mathrm{a})}$ and difference degree $N_Q^{(\mathrm{d})}$ of the base clustering subset Q will be improved, indicating that selecting some base clusters with high accuracy and large difference degree for integration can obtain better clustering results.

3.4 Consistent ensemble technology

Effectively integrating the base clusters is the key to improving the effect of clustering ensemble. Clustering ensemble methods include relationship matrix based method, graph based method, feature based method, etc.[20]. The clustering ensemble algorithm based on the co-association matrix[20] is effective and widely used, and it can solve the problem of cluster label correspondence of the base clustering results. The basic idea of this algorithm is to integrate the results of H' base clusters through the co-association matrix, and then use the agglomerative hierarchical clustering algorithm to obtain the clustering ensemble results. The co-association matrix $\boldsymbol{CA}$ is written as

$$\boldsymbol{CA} = \frac{1}{H'} \sum_{i=1}^{H'} \boldsymbol{M}_i(j,k) \tag{9}$$

Suppose that the clustering results P_1 and P_2 of the two base clusters are [1, 1, 2, 2, 3] and [1, 2, 1, 2, 2], respectively, according to Eqs. (4) and (9), we have

$$\boldsymbol{M}_1 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \boldsymbol{M}_2 = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{bmatrix},$$

$$\boldsymbol{CA} = \begin{bmatrix} 1 & 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1 & 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1 & 1/2 & 0 \\ 0 & 1/2 & 1/2 & 1 & 1/2 \\ 0 & 1/2 & 0 & 1/2 & 1 \end{bmatrix}.$$

The co-association matrix reorganizes the H' base clustering results into new patterns, and each element represents the probability that they are divided into the same cluster, accurately and quantitatively reflecting the similarity between the samples.

The agglomerative hierarchical clustering algorithm[5] adopts the idea of "bottom-up" clustering. First, each sample in the data set is regarded as a separate cluster, and the similarity is determined by calculating the distance between the two clusters. The shorter the distance is, the higher the similarity is. Then, the two most similar clusters are combined step by step and iterated repeatedly until the termination condition is reached or finally classified into one cluster, as shown in Fig.3. For six clusters, it means that each sample is one cluster. For three clusters, they are *AB*, *CDE* and *F*, respectively. For one cluster, it is *ABCDEF*. According to the cluster tree, the specific situation of different group numbers can be obtained.

4 DBSCAN-based selective clustering ensemble of structural plane attitude

4.1 Base cluster generation phase

By generating different base clusters, the characteristics of the data set can be revealed from different levels, and the deficiency of a single clustering model can be addressed[21]. In this paper, the DBSCAN algorithm is used

as the clustering algorithm. Taking advantage of the fact that DBSCAN is sensitive to the two hyperparameters r and T_d, based on the structural plane attitude data set, by combining different r and T_d values, the DBSCAN clustering algorithm is repeatedly used to calculate for H times, and H different base clustering results $P = \{P_1, P_2, \cdots, P_H\}$ are obtained. It is worth mentioning that, unlike other algorithms, DBSCAN does not need to specify the number of clusters in advance, which can reduce the error caused by human subjective judgment, and the implementation is simple and convenient.

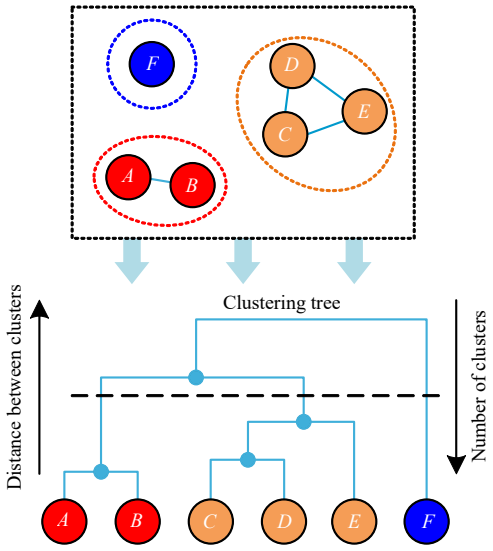


Fig. 3 Illustration of agglomerative hierarchical clustering

4.2 Base cluster selection phase

The core content of the selective clustering ensemble technology is to design appropriate selection strategies. Combining the quality and difference degree of base clustering is a relatively effective selection strategy of base clustering[14]. It not only considers the quality of base clustering, eliminates the influence of poor base clusters, but also considers the differences between base clusters, which effectively improves the clustering effect. The specific implementation steps are described as follows.

Firstly, all the base clusters are evaluated according to the performance measurement, and the base cluster with the best clustering effect is selected as the reference (the clustering result is recorded as P^o), and the accuracy of the remaining base clusters is

$$N_{p_i}^{(a)} = \frac{\left\| \boldsymbol{M}_i, \boldsymbol{M}^o \right\|}{H} \tag{10}$$

where $\boldsymbol{M}^o$ is the incidence matrix of P^o; and $\left\| \boldsymbol{M}_i, \boldsymbol{M}^o \right\|$ represents the similarity of two matrices. In this paper, the matrix is vectorized, and the vector cosine value is used to measure the similarity between the two matrices. The more similar $\boldsymbol{M}_i$ and $\boldsymbol{M}^o$ are, the better the clustering quality of P_i is.

Then, the difference degree of the base cluster is defined as

$$N_{p_i}^{(d)} = 1 - \frac{\sum_{j=1, j\neq i}^{H} \left\| \boldsymbol{M}_i, \boldsymbol{M}_j \right\|}{H(H-1)} \tag{11}$$

Obviously, the larger the values of $N_{p_i}^{(a)}$ and $N_{p_i}^{(d)}$ are, the better the quality of the corresponding base cluster is, and the greater the difference degree is. However, the relationship between the quality and the difference degree of the base cluster is complex, sometimes even contradictory. Therefore, it is necessary to find a balance between them, and the objective function[22] is defined as

$$OF(P_i) = \lambda N_{p_i}^{(a)} + (1-\lambda) N_{p_i}^{(d)} \tag{12}$$

where $\lambda (0 \leqslant \lambda \leqslant 1)$ is the balance factor, which is usually set to 0.5.

After the reference base cluster is selected, the $OF(P_i)$ value of the residual base cluster is calculated according to Eq.(12). The larger the $OF(P_i)$ value is, the better the clustering quality of P_i is, and the greater the difference degree between P_i and its residual base clustering results is. The remaining base clusters are sorted in descending order according to the value of $OF(P_i)$. Then, the first H' base clusters that make the selective ensemble model have the best performance are selected for integration.

4.3 Consistency ensemble phase

After selecting the base clusters which participate in the integration, the H' base clustering results are organized into new patterns using the co-association matrix, and the similarity between the samples is quantitatively characterized again. If two samples are divided into the same cluster in most base clusters, according to Eq.(9), the greater the corresponding element value in the co-association matrix is, the higher the probability that the two samples are divided into the same cluster is.

Then the agglomerative hierarchical clustering algorithm is used to cluster the co-association matrix. The larger the element value in the co-association matrix is, the more similar the corresponding two samples are, and the more likely they belong to the same cluster. According to the idea of agglomerative hierarchical clustering, each sample in the data set is first regarded as a separate cluster, and then the two samples with the largest element values (except the elements on the main diagonal) in the co-association matrix are gradually merged into a cluster. The two most similar clusters are gradually merged and iterated repeatedly until the termination condition is reached or finally classified into one cluster. If the cluster contains multiple samples, the average value of the co-association matrix elements of each sample in the two clusters and all samples in other clusters is calculated to determine the similarity. Finally, according to the evaluation criteria of clustering effect, the optimal partitioning number K is determined, and the structural plane noise and isolated value are eliminated

to obtain the final selective clustering ensemble result. It is worth noting that the clustering ensemble can, to a certain extent, make up for the problem that the hyperparameters of the DBSCAN algorithm are difficult to determine and have a great impact on the clustering results by integrating a certain number of good-quality base clusters.

4.4 Algorithm procedure

The basic framework of the dominant partitioning method for the structural plane attitude of rock mass using DBSCAN-based selective clustering ensemble is described as follows.

Input: Structural plane attitude data set $X_{N\times 2}$, and algorithm hyperparameters r and T_d.

Step 1: Calculate the distance matrix $D_{N\times N}$ between two structural planes in the data set according to Eqs. (1) and (2).

Step 2: Combine the hyperparameters r and T_d of the algorithm, repeat the DBSCAN algorithm for H times, and obtain H base clustering results $P=\{P_1, P_2, \cdots, P_H\}$.

Step 3: Analyze the clustering effect of each base cluster, and select the one with the best clustering effect as the reference base cluster.

Step 4: Calculate the accuracy $N_{p_i}^{(\mathrm{a})}$, difference degree $N_{p_i}^{(\mathrm{d})}$, and objective function $OF(P_i)$ of the residual base clusters according to Eqs. (10)−(12).

Step 5: Sort the remaining base clusters in descending order according to the value of $OF(P_i)$. According to this order, perform consistent ensembles one by one using Eq.(9) and agglomerative hierarchical clustering.

Step 6: According to the clustering effect, determine the number of integrated base clusters H' and the optimal partitioning number K, and eliminate the noise and isolated values.

Output: Final partitioning results of the attitude of dominant structural planes.

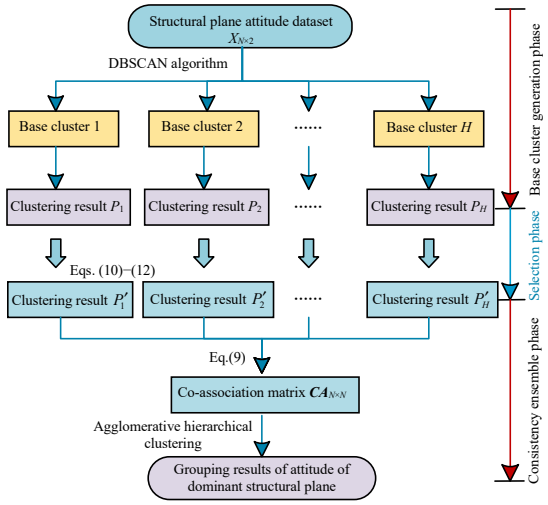


Fig. 4 Procedure of DBSCAN-based selective clustering ensemble

4.5 Clustering effect measurement

Because the clustering results have no a priori knowledge or correct clustering labels as references, the clustering quality can only be evaluated by the inherent characteristics of the data set. Since Halkidi et al.[23] proposed two clustering effect evaluation criteria, i.e. intra-cluster compact density and inter-cluster separation degree, and a series of clustering effect evaluation indicators with good performance and simple implementation was proposed. In this paper, the commonly used silhouette coefficient[9] (SC) is used as the clustering effect measurement index of structural plane data. Define the SC $s(n)$ of a sample x_n in the dataset as

$$s(n)=\frac{b(n)-a(n)}{\max[a(n),b(n)]} \tag{13}$$

where $a(n)$ is the average distance between sample x_n and other samples in the same cluster; and $b(n)$ is the average distance between sample x_n and all samples of the nearest cluster.

The average value $s(n)$ of all samples is called SC of the clustering result, and the expression is

$$\mathrm{SC}=\frac{1}{N}\sum_{n=1}^{N}s(n) \tag{14}$$

The SC value range is [−1, 1]. The larger the value is, the closer the samples in the same cluster is, the farther the samples in different clusters is, and the better the clustering effect is.

5 Algorithm verification

In order to verify the feasibility and effectiveness of the DBSCAN-based selective clustering ensemble in the dominant partitioning of structural plane attitude of rock mass, the attitude data of structural planes (195 structural planes) in the "exampmin.dips" file of DIPS software are used as the test data set, and the pole isodensity diagram of structural planes drawn by DIPS is shown in Fig.5. Based on the test data, this section mainly carries out the following two tasks: (1) Verification test: based on the DBSCAN clustering algorithm, compare the performance differences between single optimal base cluster, all integrated base cluster, and selective clustering ensemble model in the dominant partitioning of structural plane attitude, and verify the feasibility and effectiveness of the selective clustering ensemble technology; (2) Comparative test: compare the performance difference between the dominant partitioning method of structural plane attitude using DBSCAN-based selective clustering ensemble and the conventional clustering method, and test the reliability and superiority of the proposed method.

5.1 Verification test

The values of hyperparameters r and T_d of the DBSCAN algorithm are listed in Table 1. By combining different DBSCAN hyperparameters and training structural surface data set repeatedly, 24 base clustering results $P=$

$\{P_1, P_2, \cdots, P_{24}\}$ are generated. The SC of each base cluster is calculated, as shown in Table 2. It can be seen that when $r = 0.20$ and $T_d = 5$, the SC of the base cluster is the largest and the clustering effect is the best. Therefore, it is defined as the reference base cluster.

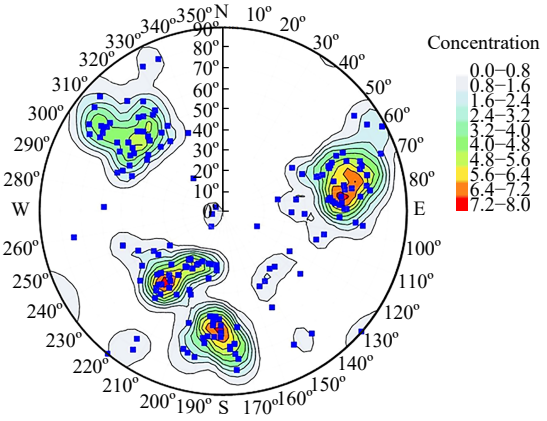


Fig. 5 Isodensity contour of the poles of structural planes

Table 1 Hyperparameters of DBSCAN algorithm

Hyperparameter	Value
r	0.10, 0.15, 0.20, 0.25, 0.30, 0.35
T_d	2, 3, 4, 5

Table 2 Silhouette coefficients of each base clusters

T_d	SC values corresponding to different r values					
	0.10	0.15	0.20	0.25	0.30	0.35
2	0.23	0.31	0.41	0.34	0.43	0.04
3	0.20	0.41	0.44	0.38	0.43	0.04
4	0.21	0.38	0.45	0.44	0.46	0.12
5	0.23	0.42	0.50	0.44	0.45	0.12

Calculate the $OF(P_i)$ value of the objective function of the remaining base clusters according to Eqs. (10)−(12), and sort it in descending order according to the $OF(P_i)$ value. Based on the reference base cluster, the base clusters are integrated one by one to form the clustering ensemble models corresponding to different numbers of base cluster. In the hierarchical clustering process, the SC values obtained by setting different group numbers K are shown in Fig.6. It can be seen that, firstly, the SC generally increases with the increasing number of base clusters participating in the integration, indicating that the clustering ensemble technology can improve the clustering effect to a certain extent. Secondly, except $K = 6$, the clustering effect of selective clustering ensemble is better than integrating all base clusters. Finally, when the number of groups $K = 5$, the clustering effect is significantly better than other groups on the whole, indicating that the data set should be divided into five groups.

Based on the clustering ensemble model with $K = 5$, it can be seen from Fig.7 that the SC corresponding to the single base cluster with the best clustering effect is 0.50, and when the first 17 base clusters are integrated, the SC increases to 0.52, which is higher than the optimal base cluster and all base clusters based ensemble model (SC = 0.51). The selective clustering ensemble algorithm is quantitatively verified in terms of performance measurement index. However, when the first 22 base clusters are integrated, the performance of the cluster ensemble model drops sharply due to the poor quality of the 21st and 22nd base clusters, which shows that some poor base clusters will affect the clustering effect.

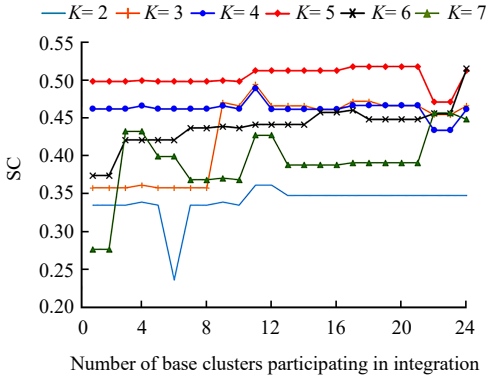


Fig. 6 Clustering effect of different partitioning numbers

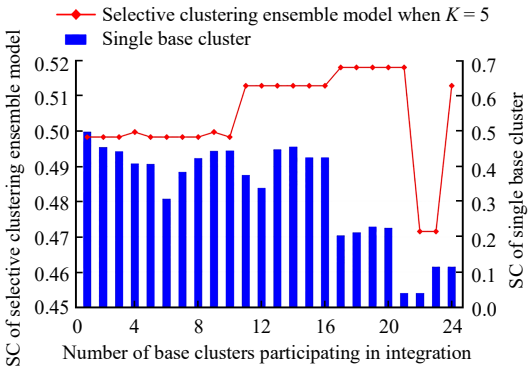
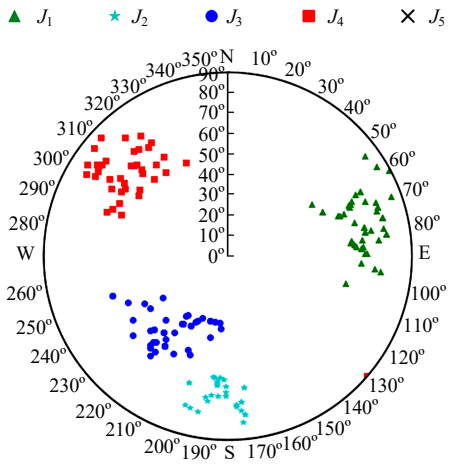
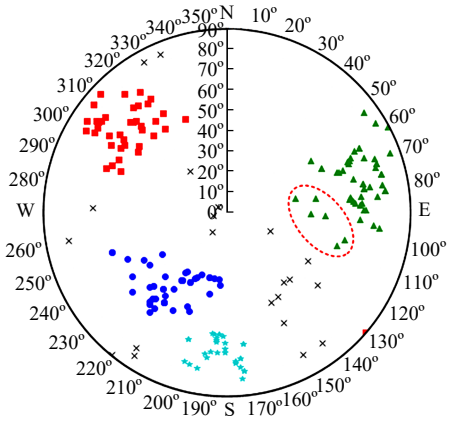


Fig. 7 Influence of the number of base clusters on selective ensemble model

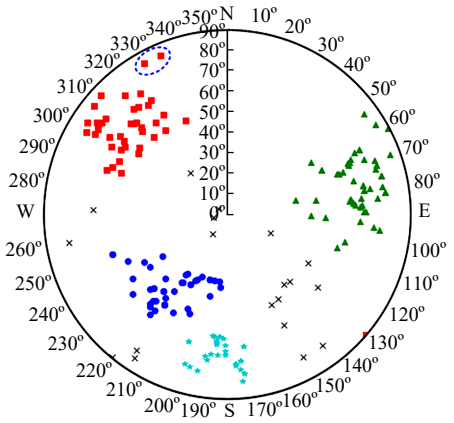
The structural plane partitioning results obtained by the optimal base cluster, full clustering ensemble model, and selective clustering ensemble model are shown in Fig.8. It can be seen that the optimal base cluster and the clustering ensemble model divide 195 structural planes into four groups and five groups, respectively. Since the noise and isolated values recognized by the single DBSCAN clustering algorithm are eliminated, for the optimal base cluster, the structural planes are divided into four groups, and the clustering results will be clustered again according to the probability that the samples of the base cluster appear in the same group. Therefore, the clustering ensemble can put the noise identified by the base cluster and the structural planes with low probability to be included in the same group into one group, as the structural plane group J_5 shown in Figs. 8(b) and 8(c). The distribution of structural plane data is considerably scattered, which makes it easy to be identified as the noise and isolated value group. Compared with Fig.8(a), it can be seen that the dominant group J_1 in the optimal base cluster has missed some structural planes, while the



(a) Optimal base cluster



(b) Full clustering ensemble



(c) Selective clustering ensemble

Fig. 8 Clustering results of the best base cluster and clustering ensemble model

clustering ensemble model can ensure that no data are missing, and accurately distinguish the dominant structural plane group from the noise and isolated value group, making up for the mis-selection or wrong selection in the single DBSCAN clustering model. Besides, the selective clustering ensemble model divides the two structural planes adjacent to J_4 into the dominant structural plane group J_4. Compared with Fig.5, it can be seen that the clustering result of the structural planes is consistent with the isodensity diagram, and also a better clustering effect is obtained, which makes the structural plane data in the group more compact while the inter-group structural plane data more separated. More reasonable and reliable clustering results can be obtained by the selective clustering ensemble model than the full clustering ensemble.

5.2 Comparative test

In order to intuitively demonstrate the advantages of the proposed selective clustering ensemble technology using the DBSCAN in the dominant partitioning of structural plane attitude, three common structural plane clustering algorithms are employed for comparative analysis in this section, i.e. *K*-means, agglomerative hierarchical clustering, and spectral clustering algorithms. In *K*-means[24], firstly, *K* points in the data set are randomly selected as the initial clustering center, and the samples are classified into groups according to the distance between the samples and the midpoint of the initial clustering. Then, the gravity center of each group is calculated as the new clustering midpoint until the clustering center of each group is no longer changed. Agglomerative hierarchical clustering[5] takes each sample as a single cluster, and gradually merges the two nearest clusters until the termination condition is reached or finally classified into one cluster. The average distance method is used to measure the distance between clusters, i.e. to calculate the average value of the distance between every two samples in two different clusters. Spectral clustering[10] converts the samples and their similarity into undirected weighted graphs, and then based on the division criteria of graph theory, the internal similarity of the divided subgraphs is made to be maximum and the similarity between subgraphs minimum. Since spectral clustering uses similarity to measure the distance between samples, $1-D(\boldsymbol{n}_i, \boldsymbol{n}_j)$ is used to measure the similarity between two structural planes. These algorithms show good clustering performance in the dominant partitioning task of structural plane, and based on different clustering strategies, which can directly reflect the effectiveness and accuracy of the DBSCAN-based selective clustering ensemble algorithm in partitioning of structural plane attitude.

Since the three algorithms all contain a crucial hyperparameter (the number of groups *K*) and the general clustering analysis needs to determine the optimal number of groups first, the number of groups $K\in[2, 7]$ is defined, and the SC values corresponding to the number of groups *K* from 2 to 7 for each algorithm are calculated, and finally, the *K* value with the best clustering effect is selected as the optimal number of groups, as shown in Fig.9. It can be seen that *K*-means, agglomerative hierarchy and spectral clustering algorithms have the largest SC when $K = 5$, 3 and 4, respectively, thus the optimal partitioning numbers corresponding to the three algorithms are $K = 5$, 3 and 4, respectively.

For the optimal grouping number, the clustering results corresponding to the three clustering algorithms are shown in Fig.10. *K*-means is suitable for the spherical data structure, thus some concentrated discrete structural

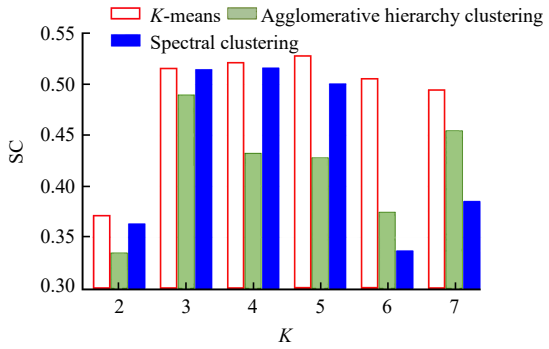
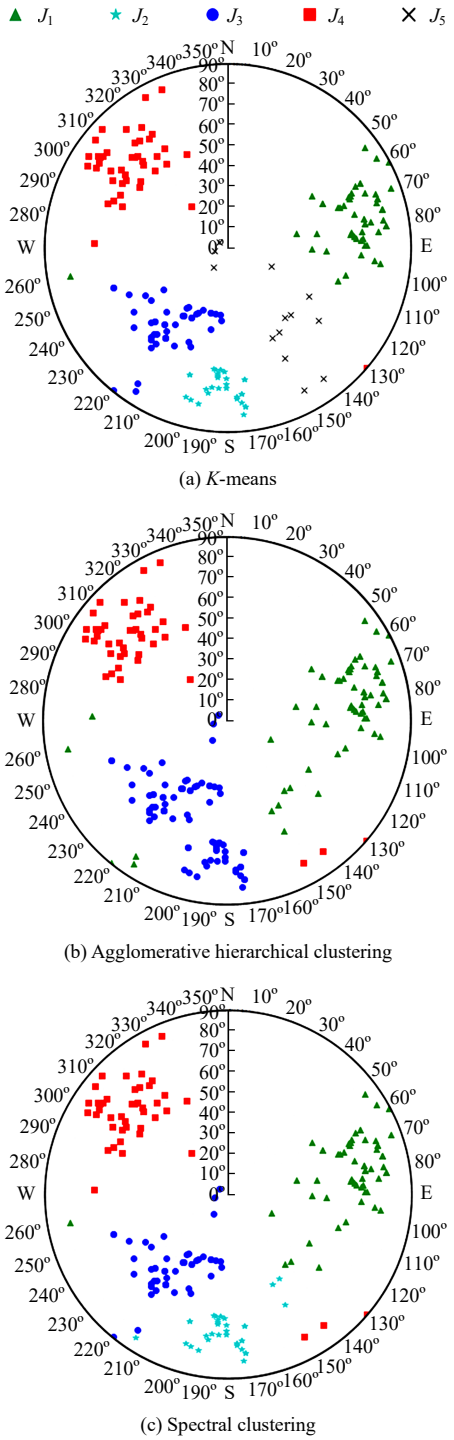


Fig. 9 Silhouette coefficients corresponding to different partitioning numbers



(a) *K*-means

(b) Agglomerative hierarchical clustering

(c) Spectral clustering

Fig. 10 Clustering results of commonly used clustering algorithms

planes are clustered into group J_5. Hierarchical clustering clusters the data into a chain structure, hence the structural planes are divided into three groups. Obviously, the clustering results are unreasonable. The boundary of the clustering results obtained by the spectral clustering algorithm is not obvious. Besides, since three algorithms do not have the function of eliminating noise and isolated values, there must be some deviation in the clustering results if all structural planes are divided into a group, which greatly affects the clustering effect. Compared with the three common clustering algorithms, the DBSCAN-based selective clustering ensemble algorithm takes the density clustering as the basic idea and couples the selective ensemble technology, which not only effectively identifies and eliminates noise and isolated values, but also makes up for the misjudgment of the single clustering model. It is superior to the conventional clustering methods. After removing noise and isolated value group, the clustering effect of the DBSCAN-based selective clustering ensemble algorithm (SC = 0.60) is significantly better than that of the three commonly used clustering algorithms, and the clustering results are more reasonable and reliable.

Compared with the commonly used single clustering method, the dominant partitioning method of structural plane attitude using the DBSCAN-based selective clustering ensemble avoids falling into local optimal solution by complementing the information of multiple base clusters, which not only makes up for the misjudgment or mis-selection of the single clustering model, but also can accurately eliminate noise and isolated values. It realizes the objective clustering analysis of structural plane attitude data. The feasibility and superiority of the proposed method in dominant partitioning of structural plane attitude are confirmed.

6 Field application

The Songta hydropower station is the first cascade hydropower station in the middle and lower reaches of Nujiang River, located in Chawalong Town, Chayu County, Tibet, China. The dam site area of the hydropower station is located in the hinterland of the Hengduan Mountains of the Qinghai-Tibet Plateau, which belongs to the landform of high mountains and valleys, and the regional structure is stable. According to the surface survey and adit logging data, the exposed strata are mainly biotite monzogranite in the late Yanshanian period, and the joints are comparatively developed, mostly with gentle and steep dip angles. Firstly, the distribution law of structural planes is analyzed to determine the dominant attitude of structural plane, and then the mechanical behavior of rock mass and engineering stability are analyzed. Three hundred and five attitude data of structural planes were collected during the investigation of an adit at the dam abutment, and the pole isodensity diagram is shown in Fig.11.

The proposed DBSCAN-based selective clustering

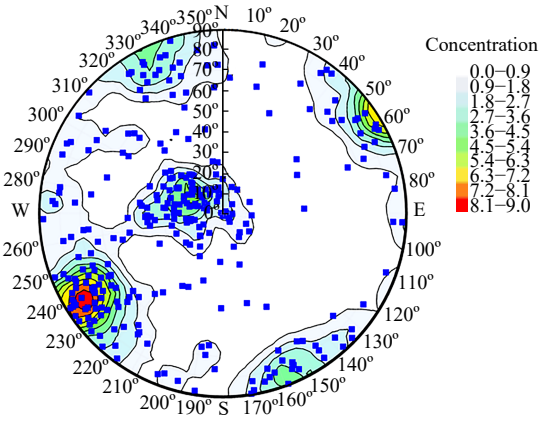


Fig. 11 Isodensity contour of the poles of structural planes

ensemble algorithm is applied to the partitioning of dominant structural plane attitude. Compared with the DIPS software data set, due to the large density of the adit structural plane data, the value of the density threshold T_d was appropriately increased when constructing the DBSCAN base cluster, as shown in Table 3. According to the algorithm procedure in section 4.4, the proposed method divides the structural planes into four groups (SC = 0.42), and the clustering results are shown in Fig.12.

Table 3 Hyperparameters of DBSCAN algorithm

Hyperparameter	Value
r	0.10, 0.15, 0.20, 0.25, 0.30, 0.35
T_d	5, 6, 7, 8

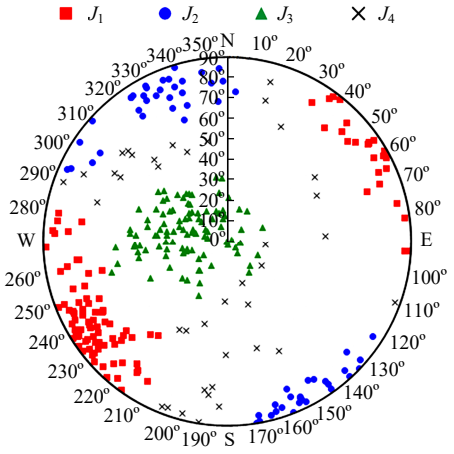


Fig. 12 Clustering result of structural planes at Songta hydropower station

It can be seen from Fig.12 that the structural planes of group J_4 are obviously a group of noise and isolated values, which should be eliminated. After excluding group J_4, the groups of dominant structural planes are shown in Fig.13. The SC is equal to 0.55, and the clustering effect is satisfactory. It can be seen that the adit has three groups of dominant structural planes, and the average attitude and number of joints of each group are listed in Table 4.

Comparing Figs. 11 and 13, it is obvious that the

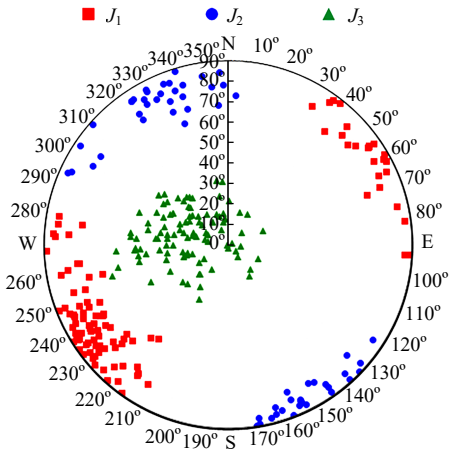


Fig. 13 Dominant attitude partitioning result of structural plane at Songta hydropower station

Table 4 Basic parameters of dominant structural plane groups

Group	Dip direction /(°)	Dip angle /(°)	Number of joints
J_1	239.55	80.29	105
J_2	332.36	82.22	59
J_3	281.59	26.32	102

partitioning results of the proposed method match the actual situation well. The proposed method can not only obtain accurate clustering results and generate a clear boundary between groups, but also effectively identify the noise and isolated values, which are difficult for the commonly used partitioning methods of dominant structural plane attitude. Meanwhile, it further demonstrates the effectiveness and advantages of the proposed method.

7 Conclusions

Based on the research results of modern artificial intelligence, the partitioning method of dominant structural plane attitude of rock mass using DBSCAN-based selective clustering ensemble is proposed, and the following conclusions are obtained:

(1) DBSCAN algorithm and selective clustering ensemble technology are introduced into the dominant partitioning of structural plane attitude. It is not necessary to specify the number of groups and the initial clustering center, which reduces the influence of subjective factors. By integrating the information of multiple base clusters, the transition from the conventional single clustering model to the ensemble model is realized, and the problem of misjudgment or mis-selection of the single model is solved.

(2) The DBSCAN-based selective clustering ensemble algorithm is used to perform the clustering analysis for the attitude data of structural planes in DIPS software. The SC of the optimal base cluster is 0.50. After eliminating the noise and isolated values, the SC reaches 0.60, which significantly improves the clustering effect and verifies the effectiveness of the selective clustering ensemble

model.

(3) Compared with three commonly used clustering algorithms, the DBSCAN-based selective clustering ensemble method can effectively identify noise and isolated values, and solve the problem that the conventional clustering methods cannot filter noise and isolated values. The clustering results of the proposed method are reasonable and reliable, and the SC is significantly higher than that of the three conventional clustering models, which shows the feasibility and advantages of the proposed method in the dominant partitioning of structural plane attitude.

(4) The proposed method is applied to the area survey of Songta hydropower station dam site. It can identify the noise and isolated values in the adit, and obtain satisfactory partitioning results that match the actual situation well, which further demonstrates the correctness and practical significance of the proposed method.

References

[1] ZHANG Jian-cong, JIANG Quan, HAO Xian-jie, et al. Analysis of stress-structural collapse mechanism of columnar jointed basalt under high stress[J]. Rock and Soil Mechanics, 2021, 42(9): 2556−2568, 2577.

[2] CAI Mei-feng, WANG Peng, ZHAO Kui, et al. Fuzzy C-means cluster analysis based on genetic algorithm for automatic identification of joint sets[J]. Chinese Journal of Rock Mechanics and Engineering, 2005, 24(3): 371−376.

[3] JIANG Shui-hua, OUYANG Su, FENG Ze-wen, et al. Reliability analysis of jointed rock slopes using updated probability distributions of structural plane parameters[J]. Rock and Soil Mechanics, 2021, 42(9): 2589−2599.

[4] GE Yun-feng, ZHONG Peng, TANG Hui-ming, et al. Intelligent measurement on geometric information of rock discontinuities based on borehole image[J]. Rock and Soil Mechanics, 2019, 40(11): 4467−4476.

[5] ZHANG Qi, WANG Qing, QUE Jin-sheng, et al. Dominant partitioning of discontinuities of rock masses based on AGNES[J]. Chinese Journal of Geotechnical Engineering, 2014, 36(8): 1432−1437.

[6] SHANLEY R J, MAHTAB M A. Delineation and analysis of clusters in orientation data[J]. Journal of the International Association for Mathematical Geology, 1976, 8(1): 9−23.

[7] HAMMAH R E, CURRAN J H. Fuzzy cluster algorithm for the automatic identification of joint sets[J]. International Journal of Rock Mechanics and Mining Sciences, 1998, 35(7): 889−905.

[8] CUI Xue-jie, YAN E-chuan, CHEN Wu. Cluster analysis of discontinuity occurrence of rock mass based on improved genetic algorithm[J]. Rock and Soil Mechanics, 2019, 40(Suppl.1): 374−380.

[9] LI X, WANG Z, PENG K, et al. Ant colony ATTA clustering algorithm of rock mass structural plane in groups[J]. Journal of Central South University, 2014, 21(2): 709−714.

[10] JIMENEZ-RODRIGUEZ R, SITAR N. A spectral method for clustering of rock discontinuity sets[J]. International Journal of Rock Mechanics and Mining Sciences, 2006, 43(7): 1052−1061.

[11] WANG W T, WU Y L, TANG C Y, et al. Adaptive density-based spatial clustering of applications with noise (DBSCAN) according to data[C]//2015 International Conference on Machine Learning and Cybernetics (ICMLC). Piscataway: IEEE, 2015: 445−451.

[12] BATTULWAR R, EMAMI E, NAGHADEHI M Z, et al. Automatic extraction of joint orientations in rock mass using PointNet and DBSCAN[C]//15th International Symposium on Visual Computing. Cham: Springer, 2020: 718−727.

[13] KUNCHEVA L I, WHITAKER C J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy[J]. Machine Learning, 2003, 51(2): 181−207.

[14] FERN X Z, LIN W. Cluster ensemble selection[J]. Statistical Analysis and Data Mining: the ASA Data Science Journal, 2008, 1(3): 128−141.

[15] SONG Teng-jiao, CHEN Jian-ping, ZHANG Wen, et al. Clustering analysis of dominative attitudes of rock mass structural plane based on firefly algorithm[J]. Journal of Northeastern University (Natural Science), 2015, 36(2): 284−288.

[16] STREHL A, GHOSH J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions[J]. Journal of Machine Learning Research, 2002, 3: 583−617.

[17] HANSEN L K, SALAMON P. Neural network ensembles[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990, 12(10): 993−1001.

[18] YANG Chun, YIN Xu-cheng, HAO Hong-wei, et al. Classifier ensemble with diversity: effectiveness analysis and ensemble optimization[J]. Acta Automatica Sinica, 2014, 40(4): 660−674.

[19] LIU Li-min. Study on clustering ensemble selection algorithm[D]. Changsha: Central South University, 2013.

[20] WANG Tong, WEI Wei, WANG Feng. Sample pairwise weighting co-association matrix based ensemble clustering algorithm[J]. Journal of Nanjing University (Natural Science), 2019, 55(4): 592−600.

[21] TOPCHY A, JAIN A K, PUNCH W. Combining multiple weak clusterings[C]//Third IEEE International Conference on Data Mining. Piscataway: IEEE, 2003: 331−338.

[22] HADJITODOROV S, KUNCHEVA L I, TODOROVA L P. Moderate diversity cluster ensembles[J]. Information Fusion Journa1, 2006, 7(3): 264−275.

[23] HALKIDI M, BATISTAKIS Y, VAZIRGIANNIS M. On clustering validation techniques[J]. Journal of Intelligent Information Systems, 2001, 17(2): 107−145.

[24] FAN Lei, WANG Liang-qing, TANG Hui-ming. Dynamic cluster analysis of discontinuity orientations of jointed rock mass[J]. Rock and Soil Mechanics, 2007, 28(11): 2405−2408.