Journal of Engineering, Project, and Production Management
2026, 16(4), 2025-287

An Efficient Unsupervised Data Mining Method Using Adaptive Grid-Entropy Density Clustering

Wei Wang[1] and Cuicui Ran[2]
[1] Lecturer, School of Information Engineering, Henan Vocational College of Agriculture, Zhengzhou, 451450, China, E-mail: WeiWangw.w@outlook.com (corresponding author).
[2] Lecturer, School of Information Engineering, Henan Agricultural Vocational College of Agriculture, Zhengzhou, 451450, China

Engineering Management
Received; revised April 29, 2026; May 8, 2026; accepted May 14, 2026
Available online May 29, 2026

Abstract: Density-based clustering algorithms in unsupervised data mining often suffer from heavy data processing loads, slow computation, and difficulties in determining density thresholds. Therefore, a modified form of clustering using an optimized algorithm for sorting point recognition is proposed herein. This algorithm initially involves adaptive partitioning of the whole space using density-based grid partitioning. This will be followed by the calculation of weighted information entropy considering the probability distribution of points in the grid. Finally, by merging adjacent high-density grids and extracting their weighted centroids, clustering analysis is performed on the centroids rather than the original data points to reduce computational complexity. The results demonstrate that the optimized algorithm reduces sample computations by over 90% in clustered datasets. When compared with mainstream density-based clustering algorithms, it achieves a response time of 19 seconds and a classification accuracy above 95% on datasets with 10,000 samples. When using a single node and running time as the benchmark, the acceleration ratio of the proposed algorithm reaches 4.6 when using two nodes. And in all five datasets, the contour coefficient remains above 0.8. From these results, it can be observed that the enhanced algorithm has a lot of advantages, such as good parallel processing capabilities, high efficiency in data mining, and high effectiveness in unsupervised learning. It can solve the problem of computational redundancy in density-based clustering algorithms.

Keywords: Adaptive grid refinement, density-based clustering, information entropy, parallel computing, unsupervised data mining.

DOI 10.32738/JEPPM-2025-287

1. Introduction

Unsupervised data mining extracts valuable knowledge and techniques from complex data sources and has attracted growing attention with the arrival of the big data era (Papakyriakou and Barbounakis, 2022). Unsupervised data mining algorithms are typically classified into clustering, association rule mining, and dimensionality reduction. As the gap between data recording and information understanding continues to grow, clustering algorithms that can operate without parameter settings have gained popularity for handling large data volumes (Chen et al., 2023). However, several limitations persist within existing clustering approaches. The K-means algorithm struggles with spatial distance metrics, hierarchical clustering is computationally expensive for big data, and density-based clustering demands precise tuning of thresholds and has a heavy computational load (Li et al., 2022). Weighted grid information entropy, a sophisticated approach that uses information entropy, utilizes weighted gridding in clustering processes to ensure optimal parameterization and data handling, leading to greater flexibility in clustering (Antunes et al., 2022). Ordering Points to Identify the Clustering Structure (OPTICS), an improved density-based clustering algorithm, does not require a predefined number of clusters. It is more suitable for unsupervised tasks and can detect clusters of arbitrary shapes with strong robustness in data recognition (Ran et al., 2023). Therefore, this paper improves OPTICS based on weighted grid information entropy and presents a novel density-based clustering algorithm. The aim is to achieve higher speed and adaptability in unsupervised data mining, offering better tool support for the coming era of digital information.

2. Related Works

Weighted grid information entropy consists of three modules: weighted grid, module similarity, and information entropy.

Information entropy reflects the probability of data occurring within system events. By applying weights to the information entropy, the grid can be assigned to different levels of importance, which helps reveal the complexity of the data. Numerous researchers have explored this topic. For instance, Wang et al. (2024) proposed a multi-scale road network based on layered road grids to address the issue of real-time updates in road networks. They used road semantics as evaluation criteria and applied information entropy to classify roads into high and low levels for matching. The results showed that their method outperformed two traditional approaches in terms of matching progress and accuracy. To determine water abundance in coal seam aquifers, Qiu et al. (2022) developed a multi-factor prediction method based on the fuzzy Delphi analytic hierarchy process and the entropy weight method. They selected various geological data sources to build a water abundance model, and the effectiveness of the method was confirmed through engineering validation. Yilahun and Hamdulla (2023) developed a method for entity extraction in knowledge graphs. They combined information entropy, information retrieval, and natural language processing into a statistical approach. By preprocessing the data, they successfully identified keywords and offered new insights into building knowledge graphs. Zhang et al. (2023) proposed a sparse sensor placement strategy based on information entropy to optimize sensor deployment in ocean environments. They used a weighted column rotation greedy algorithm for orthogonal triangle decomposition. The findings indicated a higher efficiency in cost reduction and low reconstruction error rates than traditional methods. An entropy index method was used by Tyagi and Sarma (2024) to segregate the sources of groundwater pollution using quality parameter data and present seasonally appropriate patterns. The study concluded that most samples from the tested areas belonged to the category of being hazardous based on the hazard index.

As a method to reflect data complexity, weighted grid information entropy has often been applied to improve unsupervised data mining algorithms. Many scholars have also used different approaches to optimize data mining techniques. For instance, Batool et al. (2023) proposed a prediction model for student's final grades in the field of educational data mining. This model applied random forest and artificial neural networks, using WEKA as the tool for predicting performance. Experimental results indicated that academic performance and demographic factors were the most effective attributes for prediction. To evaluate student learning outcomes, Hendrastuty et al. (2024) applied data mining through K-means clustering. They grouped students for cluster analysis and used silhouette scores as the evaluation metric. The study confirmed the effectiveness of cluster-based grouping evaluations. Hewage et al. (2023) introduced a predictive model in the field of predictive maintenance by combining K-means clustering results with convolutional neural networks. They also emphasized that major mining tasks at different stages of data mining should be updated in real time, offering new directions for predictive maintenance. To balance data mining accuracy and privacy, Tarigan et al. (2022) explored privacy-preserving data mining methods. Through the optimization process of clustering algorithms and association rule mining, they established the effectiveness of the suggested technique through 104 records and found it reliable in ensuring both precision and privacy. Nazari et al. (2024) developed an ensemble learning algorithm for predicting heart diseases. Nazari et al. (2024) also tested their theory using several datasets by assigning different weights to predict such illnesses successfully.

In summary, while many studies have advanced data mining, most enhancements to unsupervised algorithms focus on data preprocessing. There is still a gap in thoroughly exploring how to simplify complex data and set thresholds. Weighted grid information entropy effectively detects data density patterns by assigning weights to complex datasets. OPTICS, an improved density-based clustering algorithm, can group datasets of various shapes without needing parameter adjustments. This study will improve the performance of the OPTICS method through the application of weighted information entropy of grid-based clustering. This paper will provide a novel approach to data mining without supervision, with the objective of acquiring knowledge from big data.

3. Design of an Improved Unsupervised Data Mining Algorithm OPTICS

3.1. Design of Optimized Weighted Grid Information Entropy

Weighted grid information entropy assigns different weights to data based on the size of their information entropy. Data with a higher information value possesses lower entropy, whereas data with a lower information value possesses higher entropy. This partly reflects the probability of data dispersion in grid space (Turlykozhayeva et al., 2023). However, the weighing effect will be highly dependent on the selection of the unit grid division scale, hence resulting in sparse or dense grids. In this case, this research tries to optimize the technique based on the grid division logic. The basic steps of weighted grid division and distribution are shown in Fig. 1.

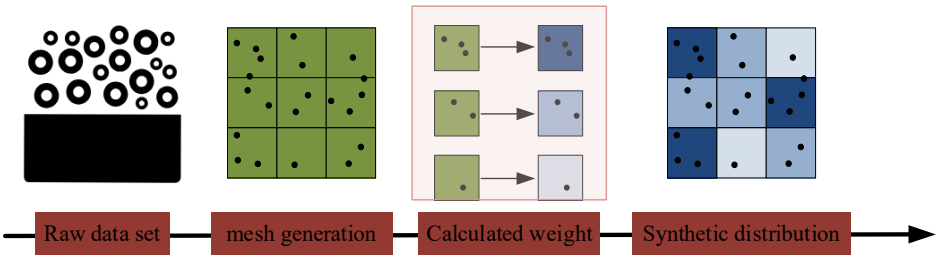


Fig. 1. The process of weighted grid information entropy partitioning and weighting

As shown in Fig. 1, weighted grid information entropy processes data through three steps: uniform grid division, grid density calculation, and weight assignment to individual grids. Weighted grid information entropy better reflects the

occurrence probability of events in the system. Many density-based clustering algorithms assign weights through grid division to effectively calculate core objects and neighborhoods, improving clustering and mining efficiency (Mvondo-She, 2023). In practice, clustering results are often affected by grid parameters. Weighted grid division typically uses uniform grids, and its calculation is shown in Eq. (1).

$$WG = (G, N(g), W) \tag{1}$$

In Eq. (1), g represents a unit grid and $N(g)$ is the set of associated cells. The information entropy of a unit grid is calculated by Eq. (2).

$$H(X) = -\sum p(x) \times lbp(x) \tag{2}$$

In Eq. (2), X represents a discrete variable, and $p(x)$ is the probability that the variable appears in system events. Uniform grid division may face issues like difficulty in determining initial parameters and imbalanced grid density (Saravanan et al., 2022). Adaptive Mesh Refinement (AMR) adjusts grid density dynamically based on the data variation. It keeps coarse grids in smooth regions and refines grids in complex areas to improve accuracy and precision (Zhang and Zhang, 2022). Therefore, this study uses AMR to improve the grid division in weighted grid information entropy. The improved process is shown in Fig. 2.

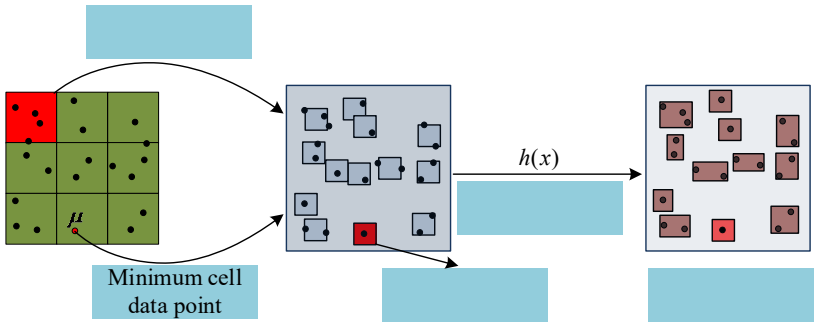


Fig. 2. Schematic diagram of AMR improved grid division logic

As shown in Fig. 2, AMR includes several parts: initial grid division, threshold calculation, and grid refinement based on the threshold. The key of the AMR algorithm is contained in its dynamic adaptation procedure. Initially, the original grid is created by considering the dimensionality of the dataset, followed by evaluation and division according to the distribution density of the data points in each grid. If the variance of the distribution of data points in any grid surpasses a pre-set limit, it will undergo further subdivision. Conversely, if the density of adjacent grid cells is lower than the merging threshold, they are merged into larger cells to reduce computational complexity. This study determines the optimal parameter combination through grid sensitivity analysis, and the initial grid size is adaptively determined based on the data dimension d to ensure efficient grid resource allocation of AMR on different datasets. The calculation of the threshold φ is shown in Eq. (3).

$$\varphi = \mu \times \min\left(\frac{1}{n}\sum_{\substack{i=1 \\ i \neq j}}^{n} \left\| p_i - p_j \right\|\right) \tag{3}$$

In Eq. (3), μ is the number of data points in the smallest unit grid. p_i and p_j represent data points in different positions. In Eq. (3), the calculation of the threshold φ relies on the statistical characteristics of the data itself, achieving automation in parameter setting and avoiding subjective presets. The function for adaptive grid size adjustment based on the threshold is shown in Eq. (4).

$$h(x) = h_0 \times (1 + \alpha \left\| \nabla \mu h(x) \right\|)^{-\beta} \tag{4}$$

In Eq. (4), $h(\bullet)$ is the function controlling local grid density. h_0 is the initial grid size determined by the data dimension. Before grid adjustment, AMR uses an error indicator to determine whether the grid needs to be refined, as shown in Eq. (5).

$$\eta K = h_k \left\| R(\mu_h) \right\| Lp(K) + b_k \tag{5}$$

In Eq. (5), h_k is the grid size is to be evaluated. $R(\mu_h)$ is the weak form residual in finite element analysis, and b_k is the boundary term. In grid-based data analysis, the connection between cells better reflects the overall features of the dataset. Therefore, this study considers the correlation of adjacent grid data and constructs weighted grids for each

partition based on neighboring grids and boundary extension. The weighted mechanism is shown in Fig. 3.

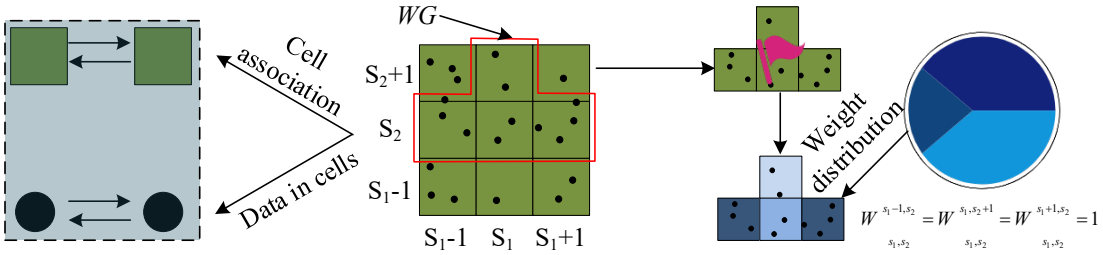


Fig. 3. Schematic diagram of the weighting process of adjacent grid groups

As shown in Fig. 3, the improvements of neighboring grid groups involve defining the weight scope and assigning grid weights. Weighted grid groups are built for each object, making the data structure more layered. Both intra-grid and inter-grid relations become important indicators for connectivity analysis. The process starts by judging whether a grid object is related to another using boundary points and determines the scope using neighboring grids. Then, the weights are assigned based on data density and the correlation of boundary points. The calculation of the weighted grid scope is shown in Eq. (6).

$$N(G)\left\{N(G_{S_1,S_2,S_d})\middle|\forall c\ s.t\quad 1\le c\le d,\quad \left|s_c{}'-s_c\right|\le 1\right\} \tag{6}$$

In Eq. (6), s_c is the c-th grid unit and G_{S_1,S_2,S_d} represents spatial grid units. Grids under the same scope are expressed as $N(G_{S_1,S_2,S_d})$. During analysis, the collection of grids in the same group is treated as a whole dataset. The weighted expression of grid set G_{S_1,S_2} is shown in Eq. (7).

$$WG=(G_{s_1,s_2},G_{s_1-1,s_2},G_{s_1+1,s_2},W_{s_1,s_2}^{s_1-1,s_2}=W_{s_1,s_2}^{s_1,s_2+1}=W_{s_1,s_2}^{s_1+1,s_2}=1) \tag{7}$$

In Eq. (7), W represents the weight. The sum of weights in a grid set is 1. Weighted grid information entropy assigns different importance to grids and their neighboring ranges. This reflects their correlation and improves the accuracy of unsupervised data classification.

3.2. OPTICS Improvement based on Weighted Grid Information Entropy

Weighted grid entropy can be an excellent technique in the optimization of unsupervised data mining algorithms, hence improving the process of mining. Classical clustering techniques are mostly dependent on parameters, and weighted grids do not help in making them more robust. OPTICS is a density-based clustering technique that can be very effective in identifying complex clusters. OPTICS clustering technique does not need parameters to be set up and produces reachability graphs, representing the result of clustering (Bhaskaran and Marappan, 2023). However, OPTICS involves complex calculations and consumes more time. Therefore, this study improves OPTICS with weighted grid information entropy to enhance its efficiency. The standard OPTICS clustering process uses reachability distances, as shown in Fig. 4.

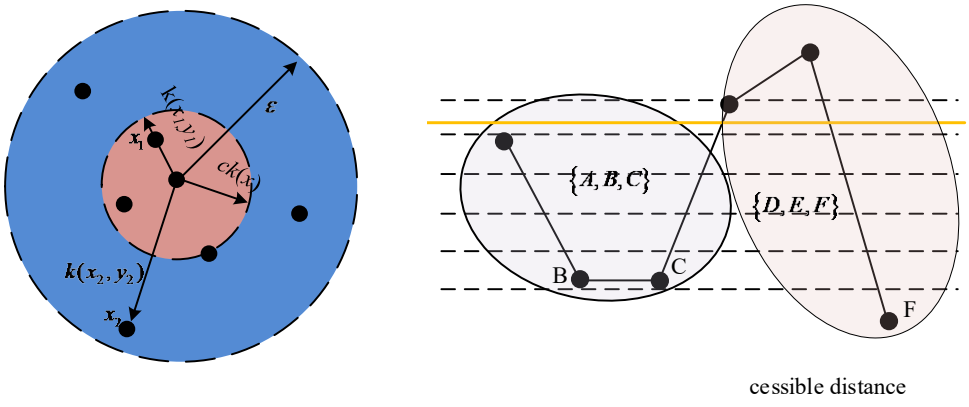


Fig. 4. Schematic diagram of the OPTICS clustering mechanism

As shown in Fig. 4, the core of OPTICS is the calculation of core distance and reachability distance. Core distance is affected by the neighborhood, and the reachability distance cannot be smaller than the core distance calculated by the neighborhood radius. Different reachability distances represent different levels of connection with the core. Adjusting the

clustering threshold affects the number of clusters. The core distance is calculated by Eq. (8).

$$ef(x)=\begin{cases}0 & N_{\varepsilon}(x)>Minpts\\ k(x,N_{\varepsilon}^{Minpts}(x)) & N_{\varepsilon}(x)>Minpts\end{cases} \tag{8}$$

In Eq. (8), the core distance is $ef(x)$. x is the sample point, ε is the neighborhood radius, $Minpts$ is the sample points in the neighborhood, and $N_{\varepsilon}(x)$ is the number of actual sample points. The reachability distance is calculated by Eq. (9).

$$rk(y,x)=\begin{cases}0 & N_{\varepsilon}(x)>Minpts\\ \max\{ck(x),k(x,y)\} & N_{\varepsilon}(x)>Minpts\end{cases} \tag{9}$$

In Eq. (9), $k(x,y)$ is the reachability distance from y to x. If this value is less than the core distance of x, then $ck(x)$ is used as the reachability distance. OPTICS automatically adjusts clustering based on data structure and is widely used in unsupervised data mining. However, its performance depends on computer power and neighborhood radius. Improved Weighted Grid Information Entropy can be used as a basis for density threshold determination and facilitates neighborhood radius calculation. This is because it does not require repetitive calculations when calculating the centroids of the weighted grids. Therefore, the proposed W-OPTICS algorithm is more effective. The key improvements are shown in Fig. 5.

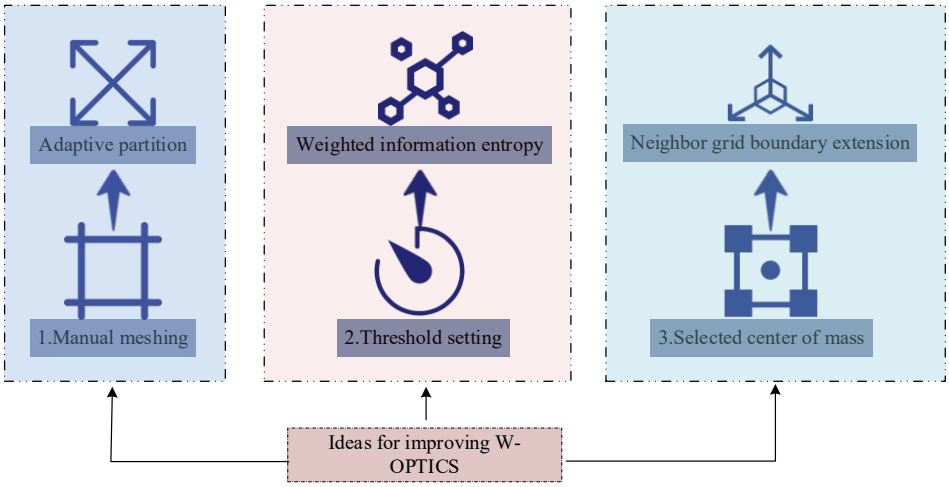


Fig. 5. Key explanation of the W-OPTICS algorithm technical improvements

From Fig. 5, it is clear that W-OPTICS enhances the methods for grid segmentation, minimum density threshold estimation, and dense point's centroid identification. Grid segmentation is done adaptively to maximize the clustering density. In addition, the enhanced method of computing the weighted grid information entropy forms the foundation of density threshold estimation. This means that W-OPTICS determines the minimum density threshold automatically using the estimated density threshold obtained from the weighted grid analysis to eliminate irrelevant grid segmentation. The cell density and minimum domain value are calculated by Eq. (10).

$$H(M)=-\sum\nolimits_{P}(m)Ibp(m) \tag{10}$$

In Eq. (10), m is the density of a grid cell after gridization. $p(m)$ is the probability of m. The weighted entropy of a grid in W-OPTICS is calculated by Eq. (11).

$$H'(M)=-\sum_{l=1}^{m}P(den(U)=a)\times IbP(den(U)=a) \tag{11}$$

In Eq. (11), $count(a)$ is the density of a. $count(l)$ is the number of divided grids. The deletion of grids not meeting the threshold is calculated by Eq. (12).

$$H''(M) = -\sum_{l=1}^{l} P(den(U) = a) \times IbP(den(U) = a) \times \frac{\sum_{l=1}^{n} \frac{a_l}{n}}{a} \tag{12}$$

In Eq. (12), $H''(M)$ is the density threshold used to delete unqualified data units. The flowchart of the proposed W-OPTICS algorithm is shown in Fig. 6.

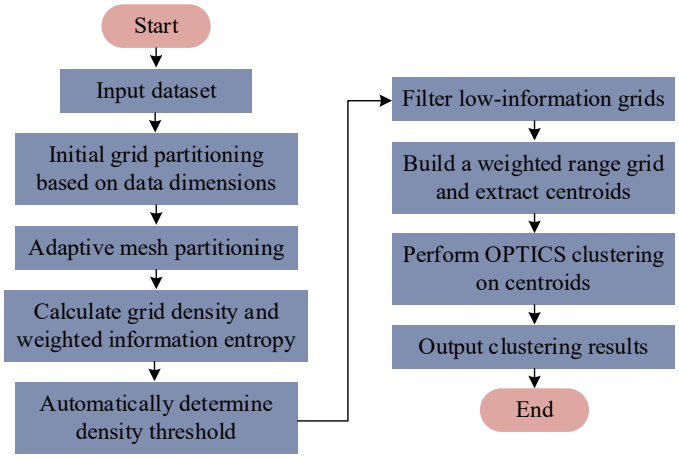


Fig. 6. Flowchart of W-OPTICS Algorithm

As illustrated in Fig. 6, the W-OPTICS algorithm first computes the density of the adaptive grid. This density will serve as an initial input parameter for clustering. Grids that are empty and grids with a density lower than the threshold value are eliminated to minimize the number of calculations required during clustering. After that, weighted range grids are employed to expand the search space. Finding centroids instead of using all data points greatly simplifies the process. The dense grid is expressed by Eq. (13).

$$B(M) = (\sum_{l=1}^{a} a_x, \sum_{l=1}^{a} a_y) \tag{13}$$

In Eq. (13), $B(M)$ is the centroid of a dense grid. The calculation is based on the weighted grid range function $N(G)$. Compared with OPTICS, W-OPTICS significantly improves the minimum density threshold calculation and simplifies the clustering process. It simplifies density-based clustering for unsupervised mining and reduces the algorithm complexity.

4. Comprehensive Performance Verification of W-OPTICS Algorithm

4.1. Simulation Analysis of W-OPTICS Performance

The artificial dataset generated by MATLAB functions and the Iris dataset provided by the UCI repository were simulated and analyzed to evaluate the performance of W-OPTICS in unsupervised data mining. Among them, the artificial datasets generated by MATLAB functions include spiral, clustering, and dense distribution. A spiral dataset is defined as a two-dimensional set of data points distributed in a spiral pattern. A clustering dataset is defined as a dataset consisting of three clusters. A dense dataset is described as a dataset with high-density region points. An example of the artificial dataset generated by MATLAB functions is shown in Table 1.

For structured datasets such as Iris and Wine, Z-score normalization was used to normalize features and eliminate the influence of dimensionality. For MNIST Digits image data, grayscale normalization and background noise filtering were performed. Geospatial data were standardized using coordinates and outlier removal to ensure consistency in data distribution. In addition, all datasets were not subjected to feature selection or dimensionality reduction to preserve their original structure for clustering analysis. The experimental environment consisted of a computer equipped with an Intel i7 processor, an NVIDIA A100 GPU, 64 GB of RAM, and an SSD, operating on Windows 10. To ensure fairness in algorithm comparison, all compared algorithms were run in the same single-threaded CPU mode based on their official implementations, without any GPU acceleration enabled. The software environment used in the experiment was MATLAB R2021a and Python 3.8, and all algorithms used their optimal default parameter configurations to avoid performance bias caused by manual parameter tuning. The clustering results of the Iris dataset using the W-OPTICS algorithm before and after processing are shown in Fig. 7.

As illustrated in Fig. 7(a), the data points of the Iris dataset appear scattered and lack clear features when displayed on a grid. In Fig. 7(b), the dataset was divided using adaptive grid partitioning. Unlike uniform distribution, this method removed grid cells below the density threshold and reduced unnecessary computation. Fig. 7(c) illustrates the centroid extraction of the dense dataset. The Iris dataset was clearly divided into three clusters located in the middle-left, upper-

right, and lower-right areas. The clusters had similar sizes and accurately captured the main features of the dataset. These results demonstrated that W-OPTICS was able to mine complex data structures effectively and precisely without requiring preset parameters. To further test the robustness and efficiency of W-OPTICS in handling multidimensional features, the centroids and runtime were evaluated. The results are shown in Fig. 8.

Table 1. Example of artificial dataset generated by MATLAB functions

Dataset	Sample ID	x-coordinate	y-coordinate	Belonging cluster (real label)
Spiral	1	0.00	0.00	-
	2	0.98	0.20	-
	3	1.62	1.14	-
Clustering	1	2.1	3.3	Cluster 1
	2	8.2	8.0	Cluster 2
	3	5.0	1.2	Cluster 3
Dense	1	4.15	4.18	-
	2	4.22	4.21	-
	3	4.18	4.19	-

Note: Iris data set is defined as UCI data set which includes 150 iris samples, 4 features, and 3 categories. An example of the Iris dataset is shown in Table 2.

Table 2. Example of Iris dataset

Sample ID	Calyx length/cm	Calyx width/cm	Petal length/cm	Petal width/cm	Category
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
3	4.7	3.2	1.3	0.2	Setosa
4	4.6	3.1	1.5	0.2	Setosa

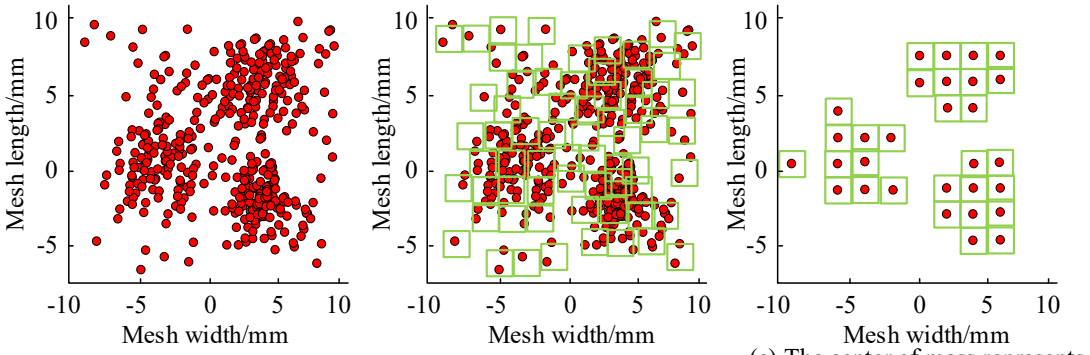


Fig. 7. Comparison of the W-OPTICS algorithm before and after clustering

As seen in Fig. 8(a), the number of data points was significantly reduced after extracting the centroids from dense grids. The spiral, clustered, and dense datasets were compressed from 1436, 3000, and 20000 points to 740, 493, and 1534 points, respectively. The point-shaped dataset was reduced by 93%, the spiral dataset by 50%, and the clustered dataset by 84%. Fig. 8(b) shows the runtime comparison before and after using centroid clustering. The initial runtimes for the original datasets were 28s, 51s, and 96s, while the runtimes after centroid clustering were reduced to 6.2s, 3.8s, and 17.3s. The dense dataset saved up to 78.7s, and the clustered dataset showed the highest efficiency improvement, with a 13.4 times speed-up. These findings revealed that W-OPTICS significantly cut down computational time while increasing overall efficiency in processing. For assessing the effectiveness of the proposed clustering algorithm, the reachability distance for the letter dataset was calculated using the OPTICS and W-OPTICS algorithms. These findings have been summarized in Fig. 9.

In Fig. 9(a), the reachability distances from OPTICS were widely scattered. Although most distances fell near 1.0mm, 1.5mm, and 3.0mm, the clusters were not clearly identifiable without further parameter tuning. In contrast, Fig. 9(b) shows

that W-OPTICS produced more concentrated reachability distances around 0.5mm, 1.0mm, and 2.0mm, revealing clearer cluster distributions. These results demonstrated that the improved W-OPTICS algorithm formed natural clusters and achieved better clustering results, making it suitable for unsupervised data mining.

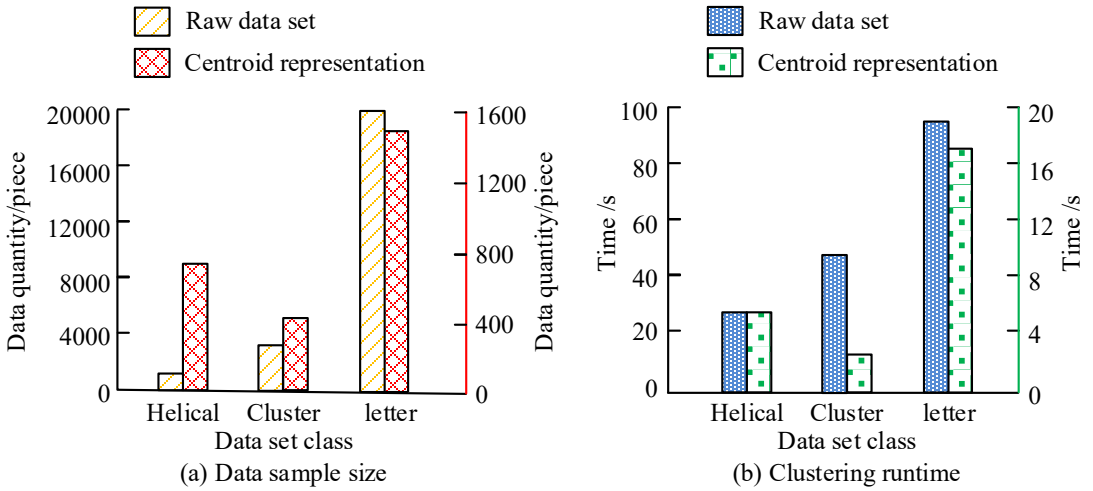


Fig. 8. Comparison of multi-dimensional data mining efficiency of the W-OPTICS

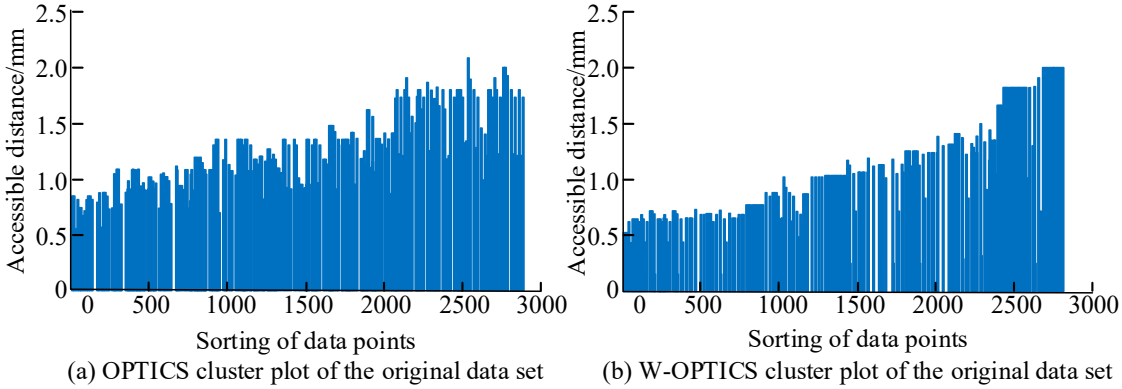


Fig. 9. Comparison of reachable distances after clustering of the two algorithms

4.2. Comprehensive Application Analysis of W-OPTICS Performance

After confirming the robustness and efficiency of W-OPTICS in unsupervised tasks, the study compared its performance with other mainstream algorithms, including HDBSCAN, Shared Nearest Neighbor Clustering (SNN), and Density-Based Clustering (DENCLUE). The comparison covered runtime, accuracy, F1 score, speedup ratio, and silhouette coefficient. Out of these, the running time denotes the amount of wall clock time that is needed by the algorithm to cluster. The accuracy means the percentage of data points that are clustered in a correct manner against the total number of data points. The value of F1 is the harmonic means of accuracy and recall, and a high value of F1 shows that the performance of the model is good. The speedup factor denotes the multiple of the running time against one node. The silhouette coefficient is used to comprehensively measure the closeness between a sample and its cluster, as well as its separation from other clusters, with values ranging from [-1, 1]. The tests were conducted on six datasets: Iris, Wine, MNIST Digits, Geospatial Data, Aggregation, and Compound, with the same computer setup mentioned earlier. The runtime and accuracy of each algorithm changed with increasing data volume are shown in Fig. 10.

As shown in Figs. 10(a) and 10(b), the runtime of all algorithms increased with the number of data points. When the data volume reached 10,000, the runtimes of W-OPTICS, DENCLUE, HDBSCAN, and SNN were 19s, 25s, 28s, and 40s, respectively. At 110,000 data points, the runtimes were 48s, 52s, 55s, and 58s. The accuracy of each algorithm decreased as the data size increased. With a volume of 10,000 data points, the clustering accuracies of W-OPTICS, DENCLUE, HDBSCAN, and SNN are 99.4%, 97.2%, 96.1%, and 90.0%, respectively. At 110,000 points, the accuracies dropped to 83.1%, 80.4%, 77.7%, and 76.2%. These results indicated that W-OPTICS maintained faster speed and higher accuracy when processing large datasets. The study also compared the clustering performance based on Receiver Operating Characteristic (ROC) curves and F1 scores, as shown in Fig. 11.

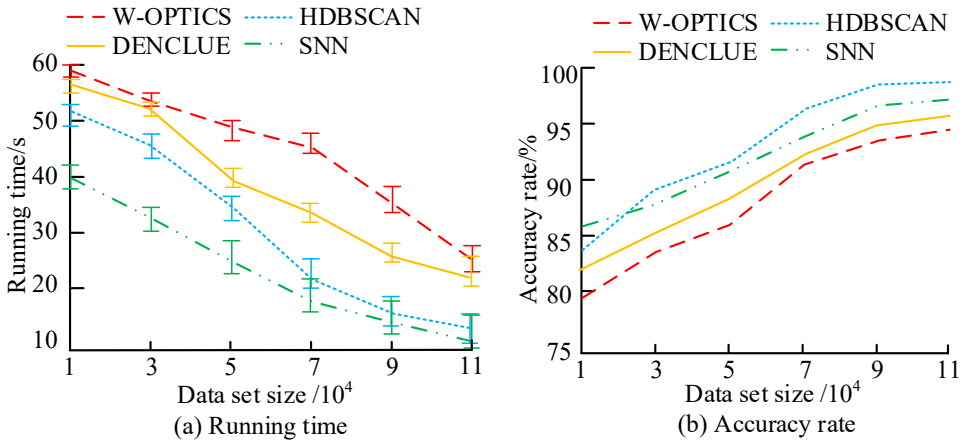


Fig. 10. Comparison of reaction time and accuracy

As shown in Figs. 10(a) and 10(b), the runtime of all algorithms increased with the number of data points. When the data volume reached 10,000, the runtimes of W-OPTICS, DENCLUE, HDBSCAN, and SNN were 19s, 25s, 28s, and 40s, respectively. At 110,000 data points, the runtimes were 48s, 52s, 55s, and 58s. The accuracy of each algorithm decreased as the data size increased. With a volume of 10,000 data points, the clustering accuracies of W-OPTICS, DENCLUE, HDBSCAN, and SNN are 99.4%, 97.2%, 96.1%, and 90.0%, respectively. At 110,000 points, the accuracies dropped to 83.1%, 80.4%, 77.7%, and 76.2%. These results indicated that W-OPTICS maintained faster speed and higher accuracy when processing large datasets. The study also compared the clustering performance based on Receiver Operating Characteristic (ROC) curves and F1 scores, as shown in Fig. 11.

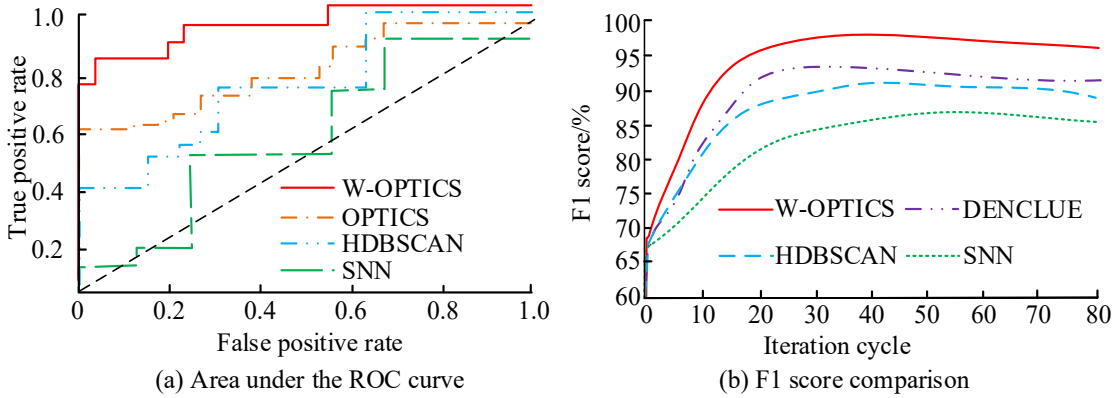


Fig. 11. ROC curve and F1 value comparison

In Fig. 11(a), the ROC curve of W-OPTICS is closer to the upper-left corner. The Area Under the Curve (AUC) values for W-OPTICS, DENCLUE, HDBSCAN, and SNN were 0.89, 0.79, 0.73, and 0.68, respectively. This result suggested that W-OPTICS had stronger classification capabilities. In Fig. 11(b), the F1 score of W-OPTICS remained above 95% after 20 iterations, significantly higher than DENCLUE's 91%, HDBSCAN's 86%, and SNN's 82%. These findings showed that W-OPTICS outperformed other algorithms in identifying true positives and in unsupervised clustering tasks. To validate the effectiveness of the W-OPTICS algorithm on real-world data, the algorithm was also applied to the Covertype dataset. In the Covertype dataset, there are 581012 data records, each with 54 attributes. It aims at predicting forest cover types from 54 attributes like terrain and soil types, which includes 7 types. The inputs will be altitude, slope, soil type, and the output will be the cover type. An example of the Covertype dataset is shown in Table 3.

Table 3. Example of Covertype dataset

Sample ID	Elevation (m)	Aspect (°)	Slope (°)	CoverType
1	2596	51	3	1
2	2590	56	2	1
3	2804	139	9	2
4	2595	45	2	2

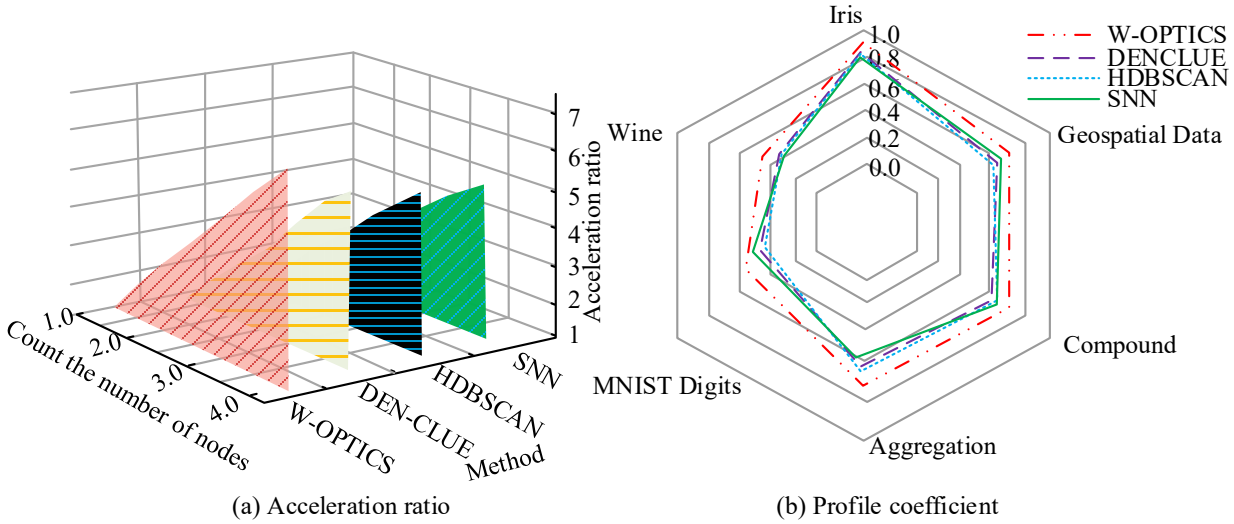
Comparative algorithms include Density Based Spatial Clustering of Applications with Noise (DBSCAN), Density

Estimation for Clustering Based on Kernel Estimation (DENCLUE), and Shared Nearest Neighbors (SNN). The experiment was conducted using a single-threaded CPU, with evaluation metrics including clustering accuracy (ACC), runtime, and normalized mutual information (NMI). The performance test results on large-scale datasets are shown in Table 4.

Table 4. Performance test results of four algorithms on large-scale datasets

Algorithm	ACC/%	Time/s	NMI
DBSCAN	63.5	285.5	0.52
DENCLUE	72.7	192.4	0.61
SNN	75.3	164.9	0.65
W-OPTICS	81.4	99.5	0.70

As shown in Table 4, even for large datasets like Covertype, the performance of the proposed W-OPTICS approach is still quite satisfactory with respect to high ACC (81.4%) and low run time (99.5 s), with 39.66% lower than the run time of the SNN clustering method. In addition, the highest value of the NMI index (0.70) also reflects its superiority in terms of structural similarity maintenance. The experimental results demonstrate that the proposed W-OPTICS algorithm can effectively reduce the computational complexity of large-scale data and has certain advantages. Speedup ratio refers to the ability to convert complex density-based clustering operations into constant-time computations through parallel processing. Under the same computational conditions, a higher speedup ratio meant better performance. The acceleration ratio experiment was conducted on a server equipped with dual Intel Xeon Gold 6248R processors (48 physical cores in total) and 512GB of memory. The algorithm is implemented based on the MATLAB parallel computing toolbox, and the density calculation and centroid extraction tasks of dense grids are allocated to multiple work nodes through parfor loops. In the experiment, the number of nodes gradually increased from one (baseline) to four to evaluate the scalability of W-OPTICS. The final experiment analyzed the speedup ratio and silhouette coefficient across different datasets, shown in Fig. 12.



(a) Acceleration ratio (b) Profile coefficient

Fig. 12. Comparison of speedup ratio and silhouette coefficient

As shown in Fig. 12(a), W-OPTICS achieved a speedup of 4.6 with 2 nodes, significantly higher than that of the other algorithms. With 4 nodes, W-OPTICS reached a speedup ratio of 6.7, again outperforming DENCLUE, HDBSCAN, and SNN, which achieved 5.4, 5.2, and 5.1, respectively. Fig. 12(b) shows that W-OPTICS achieved the highest silhouette coefficient of 0.96 on the Iris dataset and maintained higher values than the other algorithms on the remaining datasets. However, all algorithms showed low silhouette coefficients on the MNIST Digits dataset, which was likely due to the complex nature of the data. These results indicated that W-OPTICS achieved faster computation and better clustering quality across multiple datasets, showing strong adaptability in diverse applications.

5. Conclusion

To address redundant computations and unclear clustering features in the OPTICS algorithm for density-based clustering in unsupervised data mining, this study optimized OPTICS using weighted grid information entropy and proposed a new clustering algorithm, W-OPTICS. The minimum density threshold could be obtained directly during adaptive grid partitioning. By using weighted information entropy, the method identified associated dense grids and selected centroids to represent them, which reduced the overall computation. The experimental results showed that W-OPTICS reduced the sample computation by 50% for three-dimensional data using centroid-based dense grid groups. For low-dimensional data, the algorithm saved at least 83% of the sample computation. It also produced clearer features in the final reachability distance distribution. A comprehensive comparison was conducted between W-OPTICS and several mainstream density-

based unsupervised clustering algorithms. When the dataset size reached 110,000, W-OPTICS achieved a response time of 48s and a classification accuracy of 83.1%, both significantly higher than those of the other algorithms. The AUC value of W-OPTICS reached 0.89. Its F1 score remained above 95% after 20 iterations. The speedup ratio reached 6.7 when using 4 nodes. The silhouette coefficient peaked at 0.96 on the Iris dataset. Overall, W-OPTICS maintains efficient and accurate clustering performance across a wide range of datasets. It proved to be a robust unsupervised data mining algorithm suitable for density-based clustering in complex scenarios. After analyzing this article, managers will be able to rearrange their computing power and minimize their reliance on costly hardware, focusing instead on optimizing mining algorithms and deriving benefits from their efficiency. Consequently, project team members can focus less on parameter tuning and more on application scenarios. This will enable quick reaction times in cycle management by minimizing the time spent debugging clusters. Model selection prioritizes the introduction of efficient clustering schemes with automatic parameter generation to cope with data growth. In addition, in parallel deployment strategies, it is possible to plan multi-node resources earlier, fully unleash acceleration potential, and ensure decision-making timeliness.

Unlike conventional density clustering optimization techniques, which rely heavily on human intervention in selecting parameters for clustering optimization, the newly proposed W-OPTICS algorithm eliminates the need to define the clustering parameters using AMR and weighted information entropy. On the other hand, conventional techniques such as OPTICS and HDBSCAN have the limitation that their performance is dependent upon parameterization, such as neighborhood radius. As compared with K-means clustering used by Hendrastuty et al. (2024), the W-OPTICS algorithm requires no predefined cluster number. Moreover, unlike the K-means approach, the W-OPTICS algorithm can detect cluster formations of different shapes based on the density criterion. Unlike traditional OPTICS algorithms, which have limited scalability due to high computational costs, the W-OPTICS technique uses density-based grid centroid extraction to reduce computational complexity and increase scalability. The superior performance of the proposed W-OPTICS algorithm is attributed to its adaptive grid, which concentrates computing resources in high information density areas. The centroid substitution strategy shifts the algorithm complexity from relying on sample size to relying on grid density distribution, thereby significantly reducing ineffective distance calculations. Meanwhile, weighted information entropy automatically generates density thresholds, eliminating the bias and delay caused by manual trial and error. In practical deployment, the algorithm's low dependence on prior knowledge allows it to be directly embedded into automated data pipelines, making it particularly suitable for streaming processing systems that require real-time clustering. Although the proposed W-OPTICS algorithm performs well on various datasets, its effectiveness in high-dimensional composite data mining has not been validated. The algorithm is still sensitive to high-amplitude noise and outliers at distribution edges, which may affect the clarity of clustering boundaries. Therefore, future work should further explore strategies that combine W-OPTICS with feature selection or deep representation learning to optimize the processing flow for high-dimensional data by constructing low-dimensional embedding spaces. At the same time, a more robust noise recognition mechanism should be introduced, such as combining local outlier factor detection to preprocess noisy data before the grid weighting stage.

Author Contributions

Wei Wang contributed to the conception, data processing, review and revision of the article. Cuicui Ran contributed to conception, data processing and draft writing of the article.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

Declaration of Artificial Intelligence (AI) Tools

The authors used DeepSeek solely for language editing and readability improvement. The authors reviewed and verified all content and take full responsibility for the accuracy and integrity of the manuscript.

Reference

Antunes, J., Gupta, R., Mukherjee, Z., and Wanke, P. (2022). Information entropy, continuous improvement, and US energy performance: A novel stochastic-entropic analysis for ideal solutions (SEA-IS). *Annals of Operations Research*, 313(1), 289-318, DOI: 10.1007/s10479-021-04428-y.

Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H. Y., and Hussain, A. (2023). Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, 28(1), 905-971, DOI: 10.1007/s10639-022-11152-y.

Bhaskaran, S., and Marappan, R. (2023). Design and analysis of an efficient machine learning based hybrid recommendation system with enhanced density-based spatial clustering for digital e-learning applications. *Complex & Intelligent Systems*, 9(4), 3517-3533, DOI: 10.1007/s40747-021-00509-4.

Chen, C., Wu, Y., Li, J., Wang, X., Zeng, Z., Xu, J., Liu, Y., Feng, J., Chen, H., He, Y., and Xia, R. (2023). TBtools-II: A "one for all, all for one" bioinformatics platform for biological big-data mining. *Molecular Plant*, 16(11), 1733-1742, DOI: 10.1016/j.molp.2023.09.010.

Hendrastuty, N. (2024). Penerapan data mining menggunakan algoritma K-means clustering dalam evaluasi hasil pembelajaran siswa. *Jurnal Ilmiah Informatika dan Ilmu Komputer (JIMA-ILKOM)*, 3(1), 46-56, DOI: 10.58602/jima-ilkom.v3i1.26.

Hewage, U. H. W. A., Sinha, R., and Naeem, M. A. (2023). Privacy-preserving data (stream) mining techniques and their impact on data mining accuracy: A systematic literature review. *Artificial Intelligence Review*, 56(9), 10427-10464, DOI: 10.1007/s10462-023-10425-3.

Li, X. L., Chen, M. S., Wang, C. D., and Lai, J. H. (2022). Refining graph structure for incomplete multi-view clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2), 2300-2313, DOI: 10.1109/TNNLS.2022.3189763.

Mvondo-She, Y. (2023). Shannon information entropy, soliton clusters and Bose-Einstein condensation in log gravity. *Journal of High Energy Physics*, 2023(3), 1-17, DOI: 10.1007/JHEP03(2023)192.

Nazari, M., Emami, H., Rabiei, R., Hosseini, A., and Rahmatizadeh, S. (2024). Detection of cardiovascular diseases using data mining approaches: Application of an ensemble-based model. *Cognitive Computation*, 16(5), 2264–2278, DOI: 10.1007/s12559-024-10306-z.

Papakyriakou, D., and Barbounakis, I. S. (2022). Data mining methods: A review. *International Journal of Computer Applications*, 183(48), 5-19, DOI: 10.5120/ijca2022921884.

Qiu, M., Yin, X., Shi, L., Zhai, P., Gai, G., and Shao, Z. (2022). Multifactor prediction of the water richness of coal roof aquifers based on the combination weighting method and TOPSIS model: A case study in the Changcheng No. 1 coal mine. *ACS Omega*, 7(49), 44984-45003, DOI: 10.1021/acsomega.2c05297.

Ran, X., Xi, Y., Lu, Y., Wang, X., and Lu, Z. (2023). Comprehensive survey on hierarchical clustering algorithms and the recent developments. *Artificial Intelligence Review*, 56(8), 8219-8264, DOI: 10.1007/s10462-022-10366-3.

Saravanan, V., Choung, H., and Lee, S. (2022). Cell-based hybrid adaptive mesh refinement algorithm for immersed boundary method. *International Journal for Numerical Methods in Fluids*, 94(3), 272-294, DOI: 10.1002/fld.5054.

Tarigan, P. M. S., Hardinata, J. T., Qurniawan, H., Safii, M., and Winanjaya, R. (2022). Implementasi data mining menggunakan algoritma apriori dalam menentukan persediaan barang: Studi kasus Toko Sinar Harahap. *Jurnal Janitra Informatika dan Sistem Informasi*, 2(1), 9-19, DOI: 10.25008/janitra.v2i1.142.

Turlykozhayeva, D., Ussipov, N., Baigaliyeva, A., Temesheva, S., Bolysbay, A., Abrakhmatova, G., and Akhtanov, S. (2023). Routing metric and protocol for wireless mesh network based on information entropy theory. *Eurasian Physical Technical Journal*, 20(4), 46, DOI: 10.31489/2023No4/90-98.

Tyagi, S., and Sarma, K. (2024). Tracing the land use specific impacts on groundwater quality: A chemometric, information entropy WQI and health risk assessment study. *Environmental Science and Pollution Research*, 31(21), 30519-30542, DOI: 10.1007/s11356-024-33038-x.

Wang, Y., Yan, H., Li, P., and Lu, X. (2024). A multiscale road matching method based on hierarchical road meshes. *Earth Science Informatics*, 17(2), 1765-1778, DOI: 10.1007/s12145-024-01252-3.

Yilahun, H., and Hamdulla, A. (2023). Entity extraction based on the combination of information entropy and TF-IDF. *International Journal of Reasoning-based Intelligent Systems*, 15(1), 71-78, DOI: 10.1504/IJRIS.2023.128371.

Zhang, C., and Zhang, W. (2022). Efficient 2D acoustic wave finite-difference numerical simulation in strongly heterogeneous media using the adaptive mesh refinement technique. *Geophysics*, 87(1), T29-T42, DOI: 10.1190/geo2020-0801.1.

Zhang, Q., Wu, H., Mei, X., Han, D., Marino, M. D., Li, K. C., and Guo, S. (2023). A sparse sensor placement strategy based on information entropy and data reconstruction for ocean monitoring. *IEEE Internet of Things Journal*, 10(22), 19681-19694, DOI: 10.1109/JIOT.2023.3281831.

Wei Wang obtained a bachelor's of engineering degree in Computer Science and Technology from the Information Engineering University in 2006. Currently, she serves as a lecturer in the Information Security Teaching and Research Section of the School of Information Engineering at Henan Vocational College of Agriculture. She undertakes teaching and research for courses such as MySQL Database Management and Application, C Language, Fundamentals of Java Programming, and Java Programming Training. She has published more than twenty papers in several well-known domestic journals and has achieved over ten milestones so far. Her research interests include computer applications and information security.

Cuicui Ran obtained her master's degree from Nanjing University of Science and Technology in 2012. Currently, she works as a lecturer at the School of Information Engineering, Henan Agricultural Vocational College. She mainly teaches multiple courses, including Fundamentals of Big Data, Fundamentals of Java Programming, and Application of MySQL Database. She has published more than twenty papers in several well-known domestic journals and has led or participated in several research projects. Her research direction focuses on programming design.