HC3: A Three-way Clustering Method Based on Hierarchical Clustering

Wenrui Guan[a], Pingxin Wang[b,*], Wengang Jiang[a], Ying Zhang[a]

[a]*School of Automation, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu, 212100, China*
[b]*School of Science, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu 212100, China*

Abstract

Guided by the principle of three-way decision based on human-inspired computation, three-way clustering addresses the information uncertainty problem using the core region and the fringe region to characterize a cluster. The universe is split into three parts by these two regions, which capture three kinds of relationships between objects and a cluster, namely, belonging to, partially belonging to, and not belonging to. In recent years, there has been considerable three-way clustering algorithms. However, the generalization and scalability of current three-way cluster algorithms remain relatively weak, with most algorithms adhering to a fixed allocation strategy or fixed threshold parameters. In order to overcome this problem, this paper proposes a multilevel three-way clustering algorithm based on hierarchical strategy (HC3 for short). The proposed algorithm uses kernel density estimation information of data to adaptively construct a multilevel structure of data, where the higher levels (or the internal layers) with the high-density objects are closer to core regions of clusters, and the lower levels (or the external layers) withthe low-density objects are closer to fringe regions of clusters. Under the multilevel structure, we establish a three-way allocation strategy based on the stability of subclass clusters, obtaining the correct attribution of data after fully considering neighboring information. The experiments are conducted on 13 data sets with different dimensions. By comparing to other 8 clustering algorithms, the effectiveness of proposed HC3 are verified through accuracy (ACC), adjusted rand index (ARI) and adjusted mutual information (AMI).

Keywords: Three-way clustering,three-way decision theory,minimum spanning tree,uncertain data analysis.

1. Introduction

Inspired by the human cognitive processes and human intelligence, cognitive computing considers the structures and semantics of data to enhance description, prediction, and decision-making [1]. In the context of cognitive computing, clustering analysis is not only a simple data classification, but also an intelligent tool for data interpretation and knowledge discovery [2]. Specifically, clustering [3] is the task of partitioning a given unlabeled data set into multiple

*Corresponding author
Email addresses: gwr0615@163.com (Wenrui Guan), wangpingxin@just.edu.cn (Pingxin Wang), a_1_2_3@163.com (Wengang Jiang), zzzying0521@163.com (Ying Zhang)

clusters based on different criteria, such that objects within the same cluster are similar to each other while being dissimilar to objects in different clusters [4]. This allows for the discovery of the inherent data organizational structure within samples of unknown categories [5]. As an efficient technology in data mining, clustering has been widely used in various areas, including information granulation [6], machine learning [7, 8], anomaly detection [9, 10], image segmentation [11], information fusion [12], community detections [13], etc.

Traditional clustering algorithms exactly assign each object to one cluster with a crisp boundary, and this is called hard clustering [14]. However, due to the presence of information loss, such as uncertain data, noise data, outliers, etc., hard clustering fails to effectively address these challenges. Many soft computing models, such as fuzzy sets [15] and rough sets [16], have advantages in dealing with uncertain problems. Therefore, these models are combined with hard clustering algorithms to form soft clustering algorithms to deal with clustering of uncertain boundaries [17, 18]. Hard clustering and soft clustering both use a single set to depict clustering results, considering only two relationships: belonging to and not belonging to. The difference is that hard clustering does not have overlap regions while soft clustering does.

Recently, a new type of clustering algorithms [19, 20] was proposed, named three-way clustering. The foundational theory of three-way clustering is three-way decision [21, 22], which was initially used to interpret the three types of decision rules in a probabilistic rough set model [23]. Three-way decision concerns problem-solving and information processing based on a particular way of human thinking known as triadic thinking. Through expanding the decision-making results to acceptance, rejection, and delayed decisions, the three-way decision theory is more in line with human cognitions and can reduce decision error rates. Thus, three-way decision becomes a paradigm of thinking and information processing based on triadic patterns. From both macroscopic and microscopic perspectives, three-way decision can be classified into a narrow sense and a wide sense of three-way decision. The narrow sense of three-way decision mainly focuses on the theory and application of decision-theoretic rough set. The wide sense of three-way decision is built on a philosophy of thinking, problem-solving, and computing in threes. The wide sense of three-way decision offers a new understanding. By focusing on "three-way" as the use of threes, many new research areas were fostered in recent years, such as three-way classification [24], three-way formal concept analysis [25], three-way granular computing [26], three-way recommendation [27], three-way decision support [28], three-way graph convolutional [29], three-way conflict analysis [30, 31], and many others.

The triadic thinking of three-way decision offers a novel motivation and strategy for clustering data with uncertainties. Guided by three-way decision, three-way clustering introduces a unique classification strategy aimed at enhancing the algorithm's decision-making process when facing uncertain objects. Specifically, three-way clustering defines three types of membership relations between an object and a cluster, including belonging to fully, belonging to partially and not belonging to. Each cluster of three-way clustering is represented by the core and the fringe regions, and the universe is split by these two sets into three sections, which capture three kinds of relationships between objects and a cluster. Compared to the hard clustering methods, three-way clustering incorporates the fringe region to describe the uncertain relationship between objects and clusters, which provides more information about the clustering structure.

Since the introduction of three-way clustering, numerous researchers have contributed to the development of interesting and innovative approaches within this framework. These approaches can be categorized into evaluation-based three-way clustering and operator-based three-way clustering [20, 32, 33]. Evaluation-based three-way clustering algorithms employ an evaluation function and a pair of thresholds to generate three-way clustering results. The evaluation function

captures the relationship between each object and a cluster, while the thresholds determine the requirements for inclusion in distinct regions [34]. Example approaches in this category include three-way k-means [35], neighborhood rough sets based three-way clustering [36], shadowed set-based three-way clustering [37], game-theoretic rough sets based three-way clustering [38], universal gravitation based three-way clustering [39], sample similarity based three-way clustering [40]. A key issue in all these approaches is the determination of suitable thresholds. To overcome the limitations of fixed thresholds in evaluation-based three-way clustering, some adaptive three-way clustering algorithms were proposed [41, 42], which enrich the theory of 3W clustering. The operator-based three-way clustering algorithm involves the use of operators or methodologies to construct the three regions of clusters without the need for threshold determination. Examples approaches in this category include morphological operations based three-way clustering [43], image blurring and sharpening inspired three-way clustering [44] an active learning based three-way clustering using low-rank matrices [45] and multistep three-way clustering [46]. In addition to the aforementioned papers, there are other contributions [47, 48, 49, 50, 51] that enrich the theories and models of three-way clustering

Recently, Du et al. [46] proposed a multi-step three-way clustering method, which divides the universe into a multi-level structure according to human settings and focused on the data at the highest level, greatly improving the clustering accuracy. However, due to the close relationship between hyperparameters and the stability of clustering results, the hierarchical number needs to be continuously adjusted to achieve the optimal result.

In order to overcome this problem, this paper proposes a multilevel three-way clustering algorithm based on hierarchical clustering (HC3 for short). We design an erosion margin indicator to peel away the data layer by layer, allowing high-level data to converge towards the core region of clusters and adaptively obtain separation thresholds. Under the multilevel structure, we establish a three-way allocation strategy based on the stability of subclass clusters, obtaining the correct assignment of each sample after fully considering neighboring information. The main contributions of the paper can be summarized as follows:

(1) We propose a novel three-way clustering, called HC3, that utilizes an erosion margin indicator to progressively divide the universe into multi-layers and different assignment strategies are adopted for different layers.
(2) As part of the proposed algorithm, we design an erosion tolerance index by conducting statistical analysis on the data and adaptively determine the optimal number of layers for different data and perform hyperparameter optimization on the classical hierarchical model.
(3) Under the multilevel structure, we build a three-way allocation strategy based on subtree stability. The subtree stability-based three-way allocation strategy is also a non-parametric allocation model, which enriches the parameter-free allocation algorithm for three-way clustering.

The rest of this paper is formulated as follows: Section 2 introduces the background knowledge, extended definitions, and symbol meanings used in this paper. Section 3 provides a detailed description of the algorithm's process. Section 4 presents the experimental results on datasets of different sizes and types. Finally, Section 5 concludes the paper.

2. Background and related studies

2.1. Three-way representation

Let the universe $X = \{x_1, \cdots, x_n\}$ be a finite set. Suppose that the clustering divides the n objects into k clusters, labeled as $\mathbf{C} = \{C_1, C_2, \cdots, C_k\}$. Usually, a cluster can be denoted by a single set in two-way clustering. To illustrate the uncertainty of the object in the division process, three-way clustering represents a cluster C_i via two sets,

$$C_i = (Co(C_i), Fr(C_i)),$$

where, $Co(C_i)$ is the core region of C_i and $Fr(C_i)$ is the fringe region of C_i. The Universe X are divided into three parts by these two set,namely, core region $Co(C_i)$, fringe region $Fr(C_i)$, and trivial region $Tr(C_i)$. These three parts satisfy the following relations,

$$\begin{aligned}
&Co(C_i) \cap Fr(C_i) = \varnothing,\\
&Co(C_i) \cap Tr(C_i) = \varnothing,\\
&Fr(C_i) \cap Tr(C_i) = \varnothing,\\
&Tr(C_i) \cup Co(C_i) \cup Fr(C_i) = U.
\end{aligned}$$

For the samples in $Co(C_i)$, they belong to the cluster C_i and have a higher within-class similarities. For the samples in $FrC_i)$, they maybe belong to the cluster C_i and have a lower similarities with the core samples. For the samples in $Tr(C_i)$, they do not belong to the cluster C_i definitely. According to the above discussion, three-way clustering results of data set X can be represented as,

$$X \to \mathbf{C} = [C_1, \cdots, C_k] = [(Co(C_1), Fr(C_1)), \cdots, (Co(C_k), Fr(C_k))] \tag{1}$$

2.2. Kernel density estimation

Kernel Density Estimation (KDE) is a method that utilizes the distribution of data samples to derive its overall characteristics[52, 53]. Specifically, KDE utilizes a probability density function estimate of the sample to obtain data distribution properties such as data aggregation regions, by employing a probability density kernel function. The general equation for density estimation is expressed as the following equation.

$$\rho(x_i) = \sum_{x_j \in v} F\left(\frac{\|x_j - x_i\|}{h}\right), \tag{2}$$

where F is a non-negative monotonically decreasing kernel function, and the variables x_i and x_j represent sample and target points. V is a set of samples. h is the bandwidth parameter used to control the scale [54]. In this paper, the following inverse multi-quadric kernel function is employed.

$$F(x, x_i) = \sum_{x_i \in v} \frac{1}{\sqrt{\|x - x_i\|^2 + c^2}} \tag{3}$$

Traditional kernel density functions requires user-defined bandwidths, which in turn reduces adaptability. Additionally, higher density values are assigned to outlier proximity noise clusters, leading to the need for subsequent stripping operations to retain the noise. The modified kernel

density function in Equation 3 is highly localized with a bell curve shape, where the distances of the first K nearest neighbours are counted to measure the density value, and the bandwidth varies with different samples.

3. HC3: Hierarchical constrained three-way clustering

The main problems of three-way clustering is to construct the core region and the fringe region of each cluster. Inspired by the sequential three-way decision and hierarchical clustering [55], we proposes a multilevel three-way clustering algorithm based on hierarchical clustering (HC3). Firstly, we use kernel density estimation information of data to adaptively construct a multilevel structure of data, where the higher levels (or the internal layers) with the high-density objects are closer to core regions of clusters, and the lower levels (or the external layers) with the low-density objects are closer to fringe regions of clusters. After the multilevel structure is created, we establish a three-way allocation strategy, obtaining the correct assignment of each sample from high density regions to low-density regions.

3.1. Generation of hierarchical structure

Our algorithm is based on the observation that samples in high-density regions should be assigned to the cluster core, whereas samples in low-density regions should be assigned to the cluster fringe. In this work, Inspired by the concept of erosion in the CE3 algorithm, we firstly use the sample's density to adaptively construct a multilevel structure of data set. By defining erosion margin to quantitatively characterize the global density of current level, we develop a multiple iterations strategy to erode the lower density region and obtain a multilevel data structure. This hierarchical structure is conducive to improving the accuracy of clustering and the speed of the algorithm.

In order to define erosion margin, we introduce sample's neighborhood stability to reflect neighborhood distribution.

Definition 1. ***Neighborhood stability*** *Let $X = \{x_1, \cdots, x_n\}$. Suppose that the data set X has been divided $l-1$ layers $X^1, X^2, \cdots, X^{l-1}$. The remaining samples are called current layer, denoted as X^l. For the sample x_i in current layer X^l, its neighborhood stability $x_i^{(l)}$ is calculated by the variance of density of K-nearest neighbors $N_K(x_i)$, namely,*

$$x_i^{(l)} = \sqrt{\sum_{x_j \in N_K(x_i)} \frac{\|\rho^{(l)}(x_j) - \bar{\rho}^{(l)}\|^2}{K}}, \tag{4}$$

where $\rho^{(l)}(x_j)$ is the density of x_j obtained by equation (2) and $\bar{\rho}^{(l)}$ is the mean of sample's density in the K-nearest neighbors $N_K(x_i)$.

Unlike kernel density estimation, which only sums the distances between neighbors, neighborhood stability provides a more comprehensive reflection of the neighborhood distribution. A small neighborhood stability value means more concentrated neighborhood of sample x_i. In order to describe the global stability of the current layer, we give the definition of erosion margin by analyzing the data preserved in the current layer.

Definition 2. ***Erosion margin*** *Let $X = \{x_1, \cdots, x_n\}$. Suppose that the data set X has been divided $l-1$ layers $X^1, X^2, \cdots, X^{l-1}$. The remaining samples are called current layer, denoted as X^l. The erosion margin of current layer X^l is defined as follows,*

$$h(X^{(l)}) = \bar{x}^{(l)} - \sqrt{\sum_{x_i \in X^{(l)}} \frac{\|x_i^{(l)} - \bar{x}^{(l)}\|^2}{m}}, \tag{5}$$

where $x_i^{(l)}$, $\bar{x}^{(l)}$, m are the neighborhood stability of x_i , mean of neighborhood stability in X^l, the sample's number of X^l, respectively.

Erosion margin characterizes the global distribution of the current layer. A small erosion margin indicates concentrated preserved data with a relatively uniform spatial distribution. In this paper, we use erosion margin to indicate whether the current layer can continue to erode. The obtained original data set is defined as the first layer, and the density value of all samples is obtained by kernel density function. Meanwhile, the erosion margin of the samples at the highest layer is measured by equation (4). If the measured erosion margin is greater than 0, the samples at the current layer are sorted according to their density values, and the density of samples greater than the truncation density are retained to the next layer.

$$X^{(l+1)} = \left[X^{(l)} \,|\, \rho_i^{(l)} > \rho_c^{(l)} \right]. \tag{6}$$

In the above equation, $\rho_c^{(l)}$ represents the truncation density, which is set at 10% in the experiment. This means that the lower 10% of the samples in current layer are assigned to $X^{(l)}$ and the remaining 90% of the samples are considered to $X^{(l+1)}$. Repeated the above process for $X^{(l+1)}$ and update $X^{(l+1)}$ untill the following conditions are satisfied.

$$\begin{aligned} h(X^{(l)}) &> 0 \\ h(X^{(l+1)}) &\leq 0 \end{aligned} \tag{7}$$

By using the above strategy, the data set can be divided into a sequential hierarchical structure. The highest data shows distribution characteristics with low intra-class distance and high inter-class distance. The process of obtaining multilevel structure of data set X can be depicted as Algorithm 1.

Algorithm 1: Generation of hierarchical structure

Input: Data set X
Output: Hierarchical structure $(X^1, \cdots, X^L)$
for $l = 1$ **do**
  Calculate $x_i^{(l)}$ by (2) and $h(X^{(l)})$ by (5);
  Sort the samples according to their density values $x_i^{(l)}$;
  if $\rho_i^{(l)} > \rho_c^{(l)}$ **then**
    Assign x_i to X^{l+1};
  else
    Assign x_i to X^l;
  end
  Compute $h(X^{(l+1)})$ by (5)
  if $h(X^{(l+1)}) \leq 0$ **then**
    end;
  else
    $l = l + 1$
  end
end

3.2. Clustering of highest layer data

With the help of layer-by-layer erosion, a multilevel structure $(X^1, \cdots, X^L)$ is obtained. The samples at the highest layer X^L are tightly concentrated and can be used to generate original core region of each cluster. Among numerous clustering methods, the minimum spanning tree (MST for short) is a local-to-global clustering method, which is suitable for the highest layer X^L since the highest layer has small data size and large inter class differences. By using the highest level data X^L as the input and classic Prim algorithm, we obtain MST $E = [e(x_i, x_j) \mid i, j \in X^L]$[56, 57], where $e(x_i, x_j)$ is the weight of the edge between node x_i and node x_j, E is the set of all edges. In the process of clustering, the weights between intra class nodes are relatively small, while the weights between inter class nodes are significant. Therefore, we calculate the weights of all edges E and remove outliers (edges of inter class nodes) to partition the cluster. For this purpose, we designed separation coefficient s to identify outliers.

When statistically analyzing the weights of the minimum spanning tree E, it is observed that there exist significant differences between inter-class and intra-class edge weights. According to Rydder's Criterion, we set separation coefficient $s = 3Var(E)$ in this paper. By removing the weight edges greater than the separation coefficient s, the minimum tree E is divided into $d + 1$ subtrees, where d represents the number of removed weight edges. Each subtree represents an original cluster. Figure 1 provides a schematic diagram for segmenting the minimum spanning tree to obtain the original cluster. The detailed process for generating the original clusters is shown in Algorithm 2.

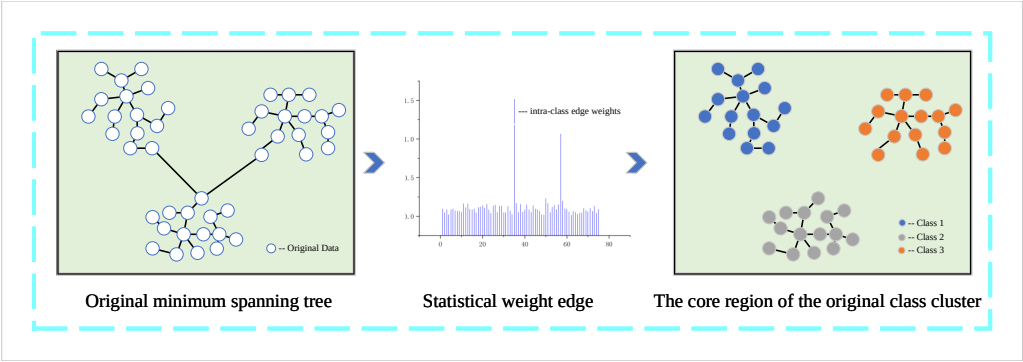


Figure 1: Segmenting the minimum spanning tree to obtain the original cluster

Algorithm 2: Clustering of Highest Layer Based on Minimum Spanning Tree

Input: X^L
Output: original core clusters $\mathbf{C} = [C_1^L, C_2^L, \cdots, C_k^L]$
Calculate the weight edge $\mathbf{D}_{n\times n}$ matrix based on X;
Create an unselected set $\mathbf{S}$; Create an selected set $\mathbf{T} = []$;
Randomly select an x_i and move x_i from set $\mathbf{S}$ to set $\mathbf{T}$;
while $\mathbf{S} \neq []$ **do**
  Calculate all connection distances between $\mathbf{S}$ and $\mathbf{T}$ and save them in the set **Dist**;
  Scan the minimum weight edge value in **Dist**,Obtain the endpoints x_i, x_j ($x_i \in \mathbf{T}, x_j \in \mathbf{S}$,) of the weighted edge;
  Move the endpoint x_j from set $\mathbf{S}$ to set $\mathbf{T}$;
  Mark the main node of x_i as x_j and store it in the matrix **Par**;
end
Obtain the minimum spanning tree E from the node relationship matrix **Par**;
Count the weighted edges in E and calculate the separation coefficient s;
Remove all weighted edges $e(x_i, x_j)$ greater than s;
Merge the remaining subtrees $E_1^L, E_2^L, \cdots, E_k^L$ into the original clusters $C_1^L, C_2^L, \cdots, C_k^L$;

3.3. Three-way allocation

The previous subsection describes the process of splitting the MST E into multiple subtrees $E_1, E_2, \cdots, E_k$ based on separation coefficient s. These original subtrees E^L (original clusters C^L) only contain a portion of the core region data. It is necessary to merge the remaining lower-level data $X^{(L-1)}, X^{(L-2)}, \cdots, X^{(1)}$ into the original clusters using a reasonable allocation strategy to complete the three-way clustering.

To achieve this, we define the distance between sample and subtree as definition 3 and propose a three-way allocation strategy based on the MST. Since the second-level data $X^{(L-1)}$ is adjacent and highly similar to the highest layer data $X^{(L)}$, the remaining data $X^{(L-2)}, X^{(L-3)}, \cdots, X^{(1)}$ is allocated in reverse order. Based on the relationships between the unallocated samples x_i and the original core region clusters, this allocation scheme assigns the remaining samples $x_i, (x_i \in X^{(L-1)}, X^{(L-2)}, \cdots, X^{(1)})$ to three types: core region belonging to a certain cluster ($x_i \in Co(C_k)$), fringe regions of some clusters ($x_i \in Fr(C_k)$), or outlier ($x_i \in outlier$),

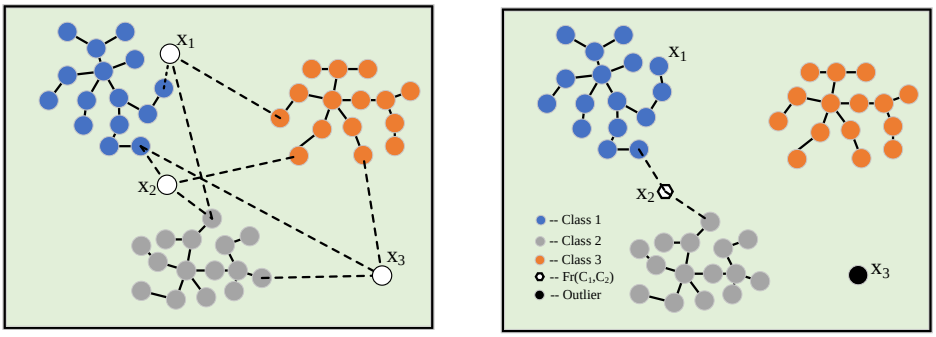
Definition 3. ***Distance between sample and subtree*** *For any unallocated sample x_i, the minimum distance between it and the subtree $E_1, \cdots, E_k$ is denoted as:*

$$psDist(C_k, x_i) = min(\|x_i - y\|), y \in C_k. \tag{8}$$

The distance between sample and subtree indicates the membership degree of the sample belonging to the cluster. Based on this point, this paper adopts the following allocation rules.

(1) If only one of the $psDist(C_p, x_i)$ is less than s, then the sample $x_i \in Co(C_p)$.
(2) If there exists more than one $psDist$ less than s, then the sample $x_i \in$ the fringe region of the cluster $(C_p, ..., C_q)$.
(3) If all $psdist$ are greater than s, then the sample $x_i \in outlier$.

As shown in figure 2, the $psDist$ of the sample x_1 is $[psDist_1, psDist_2, psDist_3]$, where only $psDist_1 < s$. Therefore, it is determined to belong to subtree E_1 after merging it into the core region of the cluster. For sample x_2, both $psDist_1, psDist_2 < s$, its membership relationship is uncertain. Therefore, the sample x_2 belongs to the fringe region of cluster C_1 and C_2. For sample x_3, all the $psDist > s$, hence it does not belong to any cluster and is classified into the outlier.



Waiting for decision allocation　　Three-way representation

Figure 2: The process of implementing three-way clustering on low layer data

Therefore, based on the calculated $psDist$, we allocate the remaining hierarchical data in reverse order, scanning and assigning layer by layer until all data points are allocated. During the clustering process, the lower-level data undergoes delayed decision-making, maximizing the allocation of the peripheral data set to the correct clusters, thereby improving the accuracy of clustering.

In summary, the overall process of the HC3 algorithm is shown in Algorithm 3.

Algorithm 3: The process of HC3 algorithm

Input: The original data $X = [x_1, \cdots, x_n]$ the size of neighborhood K.
Output: The result of three-way clustering
$\mathbf{C} = [(Co(C_1), Fr(C_1)), (Co(C_2), Fr(C_2)), \cdots, (Co(C_k), Fr(C_k))]$.
Generate multilevel structure $(X^1, \cdots, X^L)$ By Algorithm 1;
Obtain the original clusters $\mathbf{C} = [C_1, C_2, \cdots, C_k]$ by Algorithm 2 for X^L;
for $i = l - 1; i \geq 1; i - -$ **do**
  Extract unallocated layer data $X^{(i)}$;
  for $x_j \in X^{(i)}$ **do**
    Calculate $[psDist_1, psDist_2, \cdots, psDist_k]$ from the point x_j to reach clusters $[C_1, C_2, \cdots, C_k]$;
    if *only one of the* $psDist(C_p, x_i) < s$ **then**
      Assign samlple x_i to the core region of the cluster (C_p) ;
    else if *exists more than one* $psDist < s$ **then**
      Assign samlple x_i to the fringe region of the clusters $(C_p, \cdots)$;
    else if *all* $psdist > s$ **then**
      Assign samlple x_i to the *outlier* ;
    end
  end
  Update the core region of each class cluster $C_1, C_2, \cdots, C_k$;
end
Return $\mathbf{C} = [(Co(C_1), Fr(C_1)), (Co(C_2), Fr(C_2)), \cdots, (Co(C_k), Fr(C_k))]$;

The computational complexity of HC3 is dependent on the number of layers and the clusters. Assume that the data set has k clusters, and the samples are divided into L layers. The complexity of Algorithm 1 is $O(L\overline{n}^2 m_d + L\overline{n}\log\overline{n})$, where $\overline{n} = (n_1 + n)/2$, m_d is the dimension and n_1 is the number of samples in the L-th layer. The complexity of the prim algorithm is $O(n_1 + n_1 * logn_1)$. The complexity of the three-way allocation is $O((n-n_1)k)$. Therefore the overall time complexity is $O(L\overline{n}^2 m_d + L\overline{n}log\overline{n} + n_1 + n_1 * logn_1 + (n - n_1)k) \approx O(L\overline{n}^2 m_d + (L + 1)\overline{n}log\overline{n})$.

4. Experiments

4.1. Experimental description

In this section, we verify the performance of HC3 algorithm through experiments. We selected 13 data sets from the UCI Machine Learning repository [58], 7 of which are 2-dimensional data and 6 are high-dimensional data. The detailed parameters of the datasets are listed in Table 1. To compare the performances of different algorithms, eight other clustering algorithms, namely, 3W-DEPT [32], KNN-DPC [59], CE3 [43] , HDBSCAN, DBSCAN, SC, FCM, and K-means are employed in the experiments, among them, 3W-DEPT, KNN-DPC, and CE3 are three-way clustering algorithms. HDBSCAN and DBSCAN algorithms have superior performances in processing shape data and are used to compare the performances of HC3 in dealing with shape data.

To demonstrate the various performance metrics of HC3, the experimental results are presented using following indices:

(1) Accuracy (ACC). ACC is a widely used evaluation metric in clustering tasks to measure the extent to which the clustering results align with the true category markers. It quantifies the

level of agreement between the predicted clusters and the actual ground truth. In general, a higher accuracy value indicates a stronger correspondence between the clustering results and the real underlying structure of the data. ACC is an important measure for assessing the quality and effectiveness of clustering algorithms.

(2) Adjusted Rand Index (ARI). ARI is corrected based on the Rand Index. The Rand Index is used to compare the similarity of two clustering results and is defined as the number of cases in which two samples are placed in the same cluster or in different clusters. However, the Rand Index is always close to 1 for random data and therefore may overestimate the clustering effect. In order to penalise the randomness of the clustering results, ARI corrects the Rand index, taking into account the effect of a set of random clusters.

(3) Adjusted Mutual Information (AMI). AMI can adjust the degree of randomization in the evaluation adaptively and is applicable to clustering results of different sizes and densities. However, AMI has some drawbacks, such as being sensitive to the number of clustering results and having an impact on the size of the dataset.

Table 1: Data used in the experiment

	Data	Instance	Features	Class
2-dimensional data	4C	1250	2	4
	Aggregation	788	2	7
	R15	600	2	15
	Compound	399	2	6
	D31	3100	2	31
	Flame	240	2	2
	Spiral	312	2	3
real data	Ecoil	336	7	8
	Ionosphere	351	34	2
	Iris	150	4	3
	Seeds	210	7	3
	Segmantation	2310	19	7
	Wine	178	13	3

4.2. Experimental results of 2-dimensional data

We compared HC3 with 8 other clustering algorithms on 7 significant 2-dimensional data sets. The classification result figures are shown in figures 3-9, where different colors represent different clusters. In the HC3 result figures, hollow points represent data points in the fringe region, and black points represent outliers. The clustering indices of the nine algorithms are recorded in tables 2-4, where the experimental optimal value will be boldly noted.

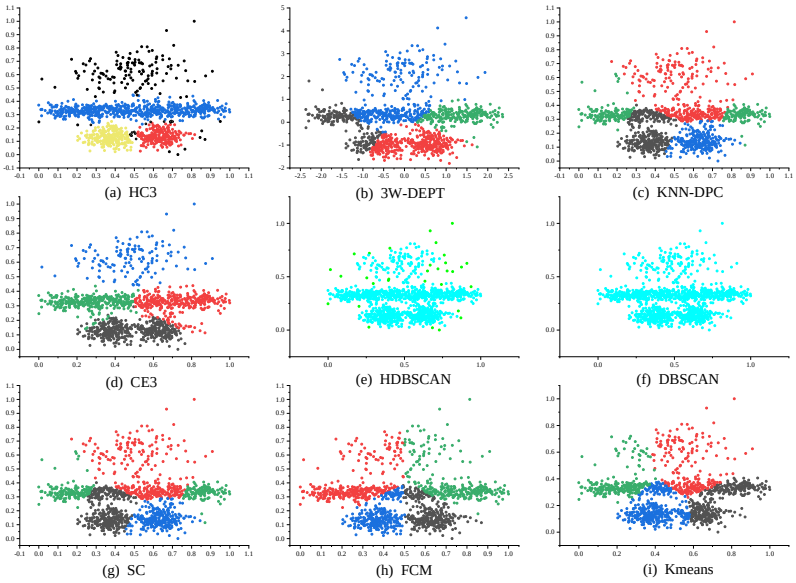


Figure 3: The clustering results of data set 4C

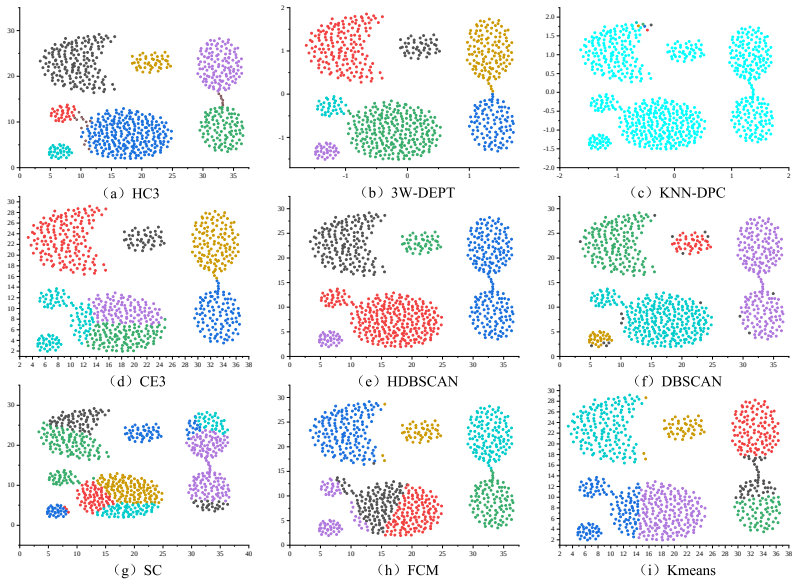


Figure 4: The clustering results of data set Aggreation

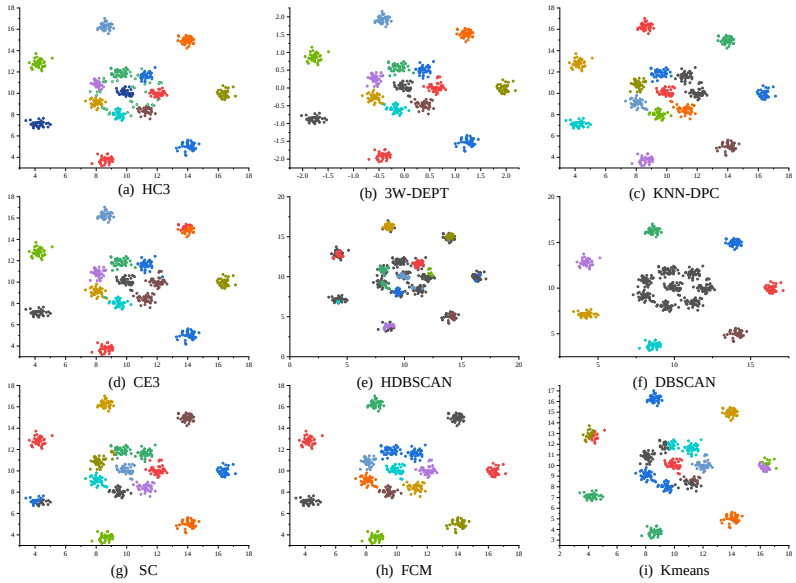


Figure 5: The clustering results of data set R15

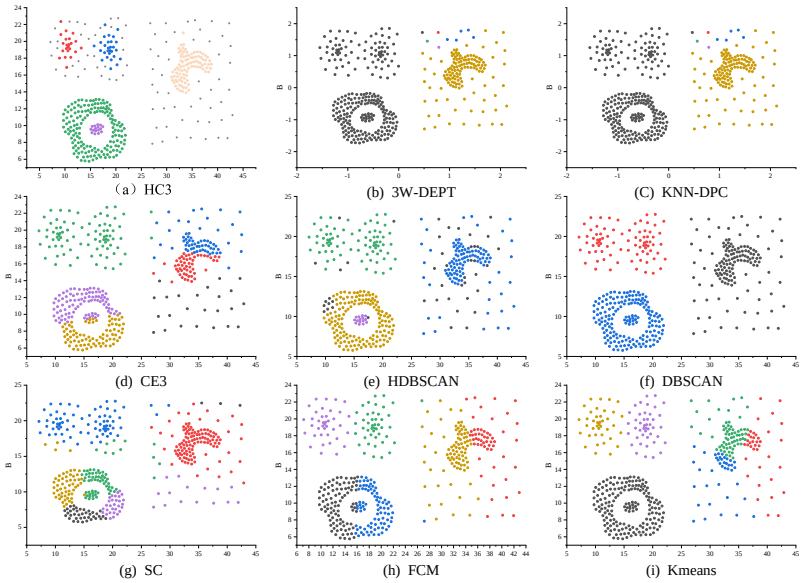
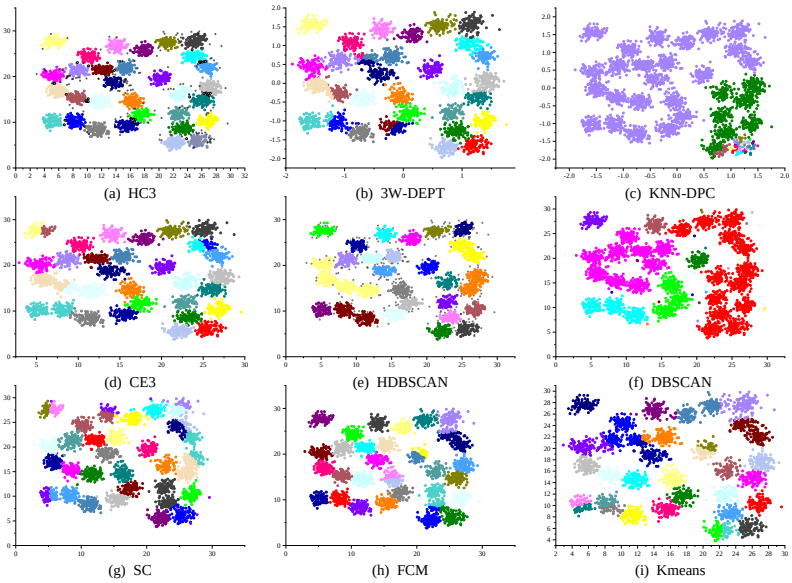


Figure 6: The clustering results of data set Compound

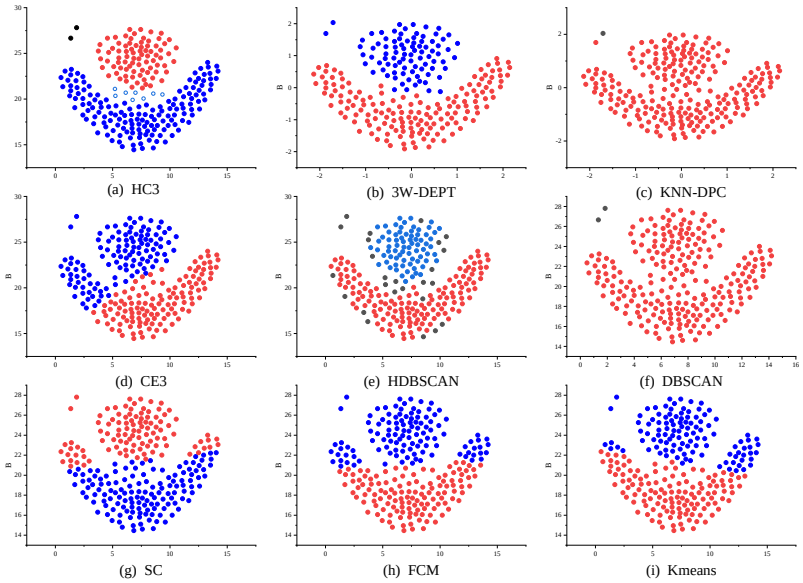


(a) HC3 (b) 3W-DEPT (c) KNN-DPC

(d) CE3 (e) HDBSCAN (f) DBSCAN

(g) SC (h) FCM (i) Kmeans

Figure 7: The clustering results of data set D31

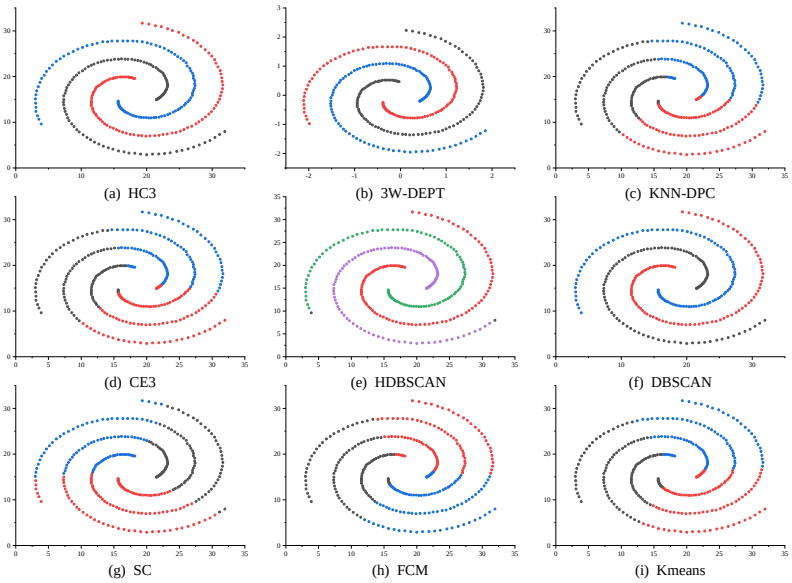


(a) HC3 (b) 3W-DEPT (c) KNN-DPC

(d) CE3 (e) HDBSCAN (f) DBSCAN

(g) SC (h) FCM (i) Kmeans

Figure 8: The clustering results of data set Flame



(a) HC3 (b) 3W-DEPT (c) KNN-DPC
(d) CE3 (e) HDBSCAN (f) DBSCAN
(g) SC (h) FCM (i) Kmeans

Figure 9: The clustering results of data set Spiral

Observing the experimental results in 2-dimensions, it can be found that the HC3 algorithm has better performance overall, and the indicators of the clustering results are overwhelmingly larger than those of other algorithms. Specific analysis for data set 4C, because of the sparse and scattered distribution of data points, compared with the tightly distributed data points below, HC3 will be put into the outlier, and the result does not affect the overall indicators. For all other data sets can be accurately classified, and the data division of the boundary region is more reasonable, so it improves the accuracy of clustering, only the AMI coefficient of D31 data set is lower than that of CE3 and 3W-DEPT algorithms. For HC3 algorithm there is only one parameter K nearest neighbour number, which has good stability, and similar results can be obtained in larger intervals, and the index is small up and down. HC3 algorithm validated for effectiveness and accuracy in 2D data sets

Table 2: Experimental results of ACC on 2D data sets.

	HC3	3W-DEPT	CE3	KNN-DPC	SC	DBSCAN	HDBSCAN	FCM	Kmeans
4C	**0.974**	0.562	0.672	0.421	0.470	0.485	0.595	0.646	0.664
Aggregation	**1.000**	**1.000**	0.758	0.347	0.807	0.820	0.813	0.794	0.802
R15	**1.000**	0.993	0.906	0.980	0.903	0.533	0.575	0.996	0.721
Compound	**1.000**	0.890	0.789	0.639	0.526	0.739	0.782	0.656	0.779
D31	**0.982**	0.900	0.895	0.075	0.866	0.229	0.589	0.891	0.800
Flame	**1.000**	**1.000**	0.825	0.641	0.821	0.645	0.895	0.850	0.858
Spiral	**1.000**	**1.000**	0.346	0.389	0.349	**1.000**	0.995	0.339	0.346

Table 3: Experimental results of ARI on 2D data sets.

	HC3	3W-DEPT	CE3	KNN-DPC	SC	DBSCAN	HDBSCAN	FCM	Kmeans
4C	**0.940**	0.462	0.582	0.084	0.404	0.000	0.381	0.491	0.263
Aggregation	**1.000**	**1.000**	0.706	0.000	0.423	0.809	0.794	0.745	0.748
R15	**1.000**	0.993	0.911	0.898	0.857	0.264	0.201	0.472	0.993
Compound	**1.000**	0.850	0.440	0.440	0.401	0.742	0.759	0.538	0.400
D31	**0.882**	0.859	**0.882**	0.038	0.791	0.155	0.579	0.428	0.877
Flame	**1.000**	**1.000**	0.408	0.000	0.534	0.007	0.810	0.486	0.463
Spiral	**1.000**	**1.000**	0.006	0.006	0.006	**1.000**	0.995	0.006	0.006

Table 4: Experimental results of AMI on 2D data sets.

	HC3	3W-DEPT	CE3	KNN-DPC	SC	DBSCAN	HDBSCAN	FCM	Kmeans
4C	**0.895**	0.610	0.707	0.210	0.551	0.000	0.497	0.507	0.400
Aggregation	**1.000**	**1.000**	0.849	0.000	0.636	0.888	0.853	0.853	0.850
R15	**1.000**	0.994	0.970	0.955	0.892	0.732	0.601	0.725	0.994
Compound	**1.000**	0.858	0.641	0.641	0.564	0.805	0.745	0.711	0.605
D31	0.875	0.925	**0.950**	0.288	0.901	0.599	0.823	0.761	0.943
Flame	**1.000**	**1.000**	0.366	0.000	0.444	0.005	0.748	0.438	0.450
Spiral	**1.000**	**1.000**	0.005	0.005	0.005	**1**.000	0.992	0.006	0.005

4.3. Experimental results of real data sets

We select six real data sets for our experiment and compared the results using the same methodology. The various experimental metrics are presented in tables 5-7, where the experimental optimal value will be boldly noted.

Table 5: Experimental results of ACC on real data sets.

	HC3	3W-DEPT	CE3	KNN-DPC	SC	DBSCAN	HDBSCAN	FCM	Kmeans
Ecoil	**0.761**	0.726	0.573	0.452	0.479	0.426	0.384	0.497	0.637
Ionosphere	**0.870**	0.641	0.712	0.638	0.818	0.641	0.769	0.709	0.709
Iris	**0.901**	0.667	0.833	0.900	0.820	0.333	0.333	0.893	0.887
Seeds	**0.968**	0.924	0.925	0.886	0.881	0.333	0.348	0.895	0.895
Segmantation	0.577	0.643	**0.655**	0.425	0.487	0.393	0.297	0.607	0.498
Wine	0.667	**0.966**	0.719	0.730	0.708	0.500	0.427	0.685	0.702

Table 6: Experimental results of AMI on real data sets.

	HC3	3W-DEPT	CE3	KNN-DPC	SC	DBSCAN	HDBSCAN	FCM	Kmeans
Ecoil	**0.688**	0.626	0.576	0.118	0.456	0.042	0.003	0.532	0.580
Ionosphere	**0.633**	0.405	0.132	0.030	0.387	0.004	0.233	0.127	0.129
Iris	**0.850**	0.729	0.653	0.773	0.640	0.015	0.015	0.745	0.737
Seeds	**0.944**	0.744	0.735	0.695	0.646	0.060	0.012	0.691	0.691
Segmantation	**0.710**	0.260	0.552	0.399	0.470	0.488	0.342	0.514	0.422
Wine	0.466	**0.875**	0.453	0.435	0.406	0.308	0.035	0.408	0.420

Table 7: Experimental results of ARI on real data sets.

	HC3	3W-DEPT	CE3	KNN-DPC	SC	DBSCAN	HDBSCAN	FCM	Kmeans
Ecoil	0.519	**0.623**	0.431	0.041	0.298	0.001	0.0013	0.368	0.467
Ionosphere	**0.523**	0.361	0.176	0.005	0.381	0.293	0.293	0.171	0.171
Iris	**0.861**	0.563	0.615	0.741	0.595	0.050	0.004	0.726	0.713
Seeds	**0.860**	0.789	0.785	0.702	0.686	0.008	0.008	0.716	0.716
Segmantation	**0.418**	0.245	0.400	0.188	0.341	0.208	0.104	0.391	0.280
Wine	0.325	**0.896**	0.395	0.365	0.349	0.262	0.003	0.349	0.366

By analysing tables 5-7, it can be seen that HC3 also performs better on the real data sets, and for most data sets the indicators are greater than the other algorithms, e.g., in the AMI indicators, HC3 improves by 5 to 10 percentage points on average compared with the other algorithms. Only in the Wine data set is the overall performance worse than the 3W-DEPT algorithm, while the rest of the data performs well. The excellent clustering performance of HC3 algorithm is illustrated based on the overall experimental results.

4.4. Analysis of Parameter Sensitivity

For HC3, the algorithm involves two parameters: the number of nearest neighbors K and the multiple of variance in the separation coefficient s. We selected three datasets R15, 4C, Aggregation for parameter sensitivity analysis. The range of $K \in (2, 15)$, and the range of variance multiples $\in (2, 5)$ with a step size of 0.25. The experimental results are shown in figure 10. Overall, the algorithm is not sensitive to two parameters. For the number of neighbors K, it is well known that when K is small, the algorithm is very sensitive to local noise and outliers in the vicinity, and is easily affected by noise, leading to inaccurate density estimation. Therefore, the accuracy of the algorithm decreases at $K = 2, 3$. Comfortingly, the ideal effect of the algorithm is achieved when $K = 4$.

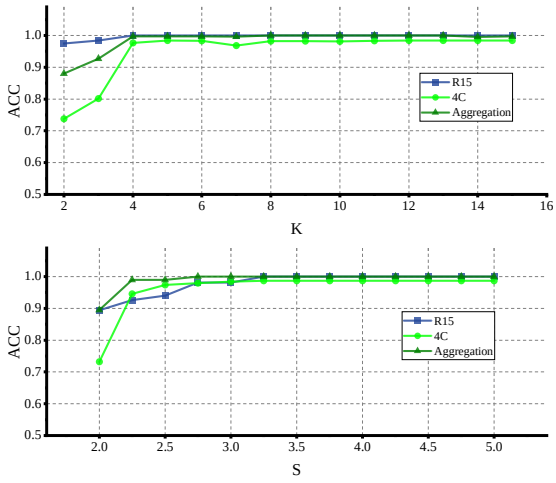


Figure 10: Analysis of Parameter Sensitivity

For the separation coefficient s, its main function is to remove outliers on the edges E of MST, which represent the edges between core clusters. Observing the experimental results, it was found that for data set 4C, when the multiple of variance is 2, the accuracy of the algorithm decreases. To investigate, we obtained the edge values of MST in data set 4C as shown in Figure 11, where the filled areas represent the range of separation coefficients. We found that when the multiple is 2, $s = 0.0198$, the algorithm erroneously removes edges within the class cluster, resulting in a decrease in accuracy. This extreme situation can be easily avoided. By observing the range of the separation coefficient, we conclude that adjusting it within an appropriate range has no effect on the algorithm.

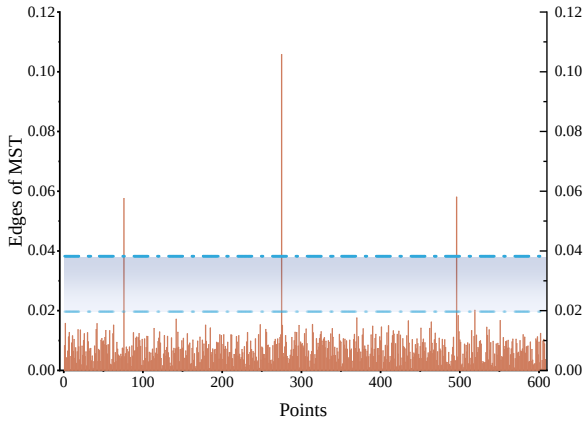


Figure 11: The Edge Values of 4C

5. Conclusions and future work

In this paper, we present a new three-way clustering algorithm that aims to address the problems of existing algorithms in terms of generalisation and scalability. Inspired by hierarchical clustering, the algorithm uses erosion margin to adaptively construct a multilevel structure of data, where the higher levels with a higher density are closer to core regions of clusters, and the lower levels with a lower density are closer to fringe regions of clusters. Under the multilevel structure, we develop a three-way allocation strategy based on the stability of subclass clusters.

To evaluate the performance of the algorithm, we conducted experiments on 13 data sets with different dimensions and compared it with eight other clustering algorithms. We used three metrics to validate the effectiveness of the algorithm. The experimental results show that our proposed algorithm HC3 outperforms the other algorithms, demonstrating its potential for practical clustering applications. This research has positive implications for advancing the field of cluster analysis and providing a more robust solution for discovering data structures in digital and information environments.

In the future work, we intend to improve the design of constructing hierarchical structure, which can can potentially enhance the efficiency and accuracy of the clustering algorithm. Determination of neighbor size parameter used in the algorithm automatically will be one of our another planned future work.

Funding This work is supported by the National Natural Science Foundation of China (nos. 62076111, 62006099) and the Key Laboratory of Oceanographic Big Data Mining & Application of Zhejiang Province (no. OBDMA202002).

Authors' Contributions Wenrui Guan: Software and writing original draft. Pingxin Wang: Supervision and methodology. Wengang Jiang: Revising the draft. Ying Zhang: Data curation and formal analysis.

Data Availability The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.
Informed Consent Informed consent was obtained from all individual participants included in the study.
Competing interests The authors declare that they have no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Yao, J., Yao, Y., Ciucci, D., Huang, K. (2022). Granular computing and three-way decisions for cognitive analytics. Cognitive Computation, 14(6), 1801-1804.
[2] Yang XB, Qi SY, Song NX, et al. (2013) Test cost sensitive multigranulation rough set: Model and minimal cost selection, Inf. Sci. 250 :184-199.
[3] Xu DK, Tian YJ (2015) A comprehensive survey of clustering algorithms, Ann. Data Sci. 2:165-193.
[4] Wu TF, Fan JC, Wang PX (2022) An improved three-way clustering based on ensemble strategy, Mathematics 10: 1457.
[5] Guo L, Zhan JM, Xu ZX, et al. (2023) A consensus measure-based three-way clustering method for fuzzy large group decision making, Inf. Sci. 632:144-163.
[6] Xu WH, Yuan KH, Li WT (2022) Dynamic updating approximations of local generalized multigranulation neighborhood rough set, Appl. Intell. 52: 9148-9173.
[7] Jordan MI, Mitchell TM (2015) Machine learning: trends, perspectives, and prospects, Science 349: 255-260 .
[8] Dong YX, Ma XJ, Fu TL (2021) Electrical load forecasting: A deep learning approach based on K-nearest neighbors, Appl. Soft Comput. 99 :106900.
[9] Liu HE, Li EH, Liu XW, et al. (2021) Anomaly detection with kernel preserving embedding, ACM Trans. Knowl. Discovery From Data, 15: 1-18.
[10] Ding WP, Nayak J, Naik B, et al. (2021) Fuzzy and real-coded chemical reaction optimization for intrusion detection in industrial big data environment, IEEE Trans. Ind. Informat. 17:4298-4307.
[11] Ding WP, Chakraborty S, Mali K, et al.(2022) An unsupervised fuzzy clustering approach for early screening of Covid-19 from radiological images, IEEE Trans. Fuzzy Syst. 30: 2902-2914,
[12] Xu WH, Yu JH (2017) A novel approach to information fusion in multi-source datasets: A granular computing viewpoint, Inf. Sci. 378 :410-423.
[13] Jiao PF, Yu W, Wang WJ, et al. (2018) Exploring temporal community structure and constant evolutionary pattern hiding in dynamic networks, Neurocomputing 314:224-233.
[14] Jain AK (2010) Data clustering: 50 years beyond k-means, Pattern Recognit. Lett. 31:651–666.
[15] Zadeh LA (1965) Fuzzy sets, Int. J. Innov Comp Inf Control. 8:338–353.
[16] Dou HL, Yang XB, Song XN, et al (2016) Decision-theoretic rough set: A multicost strategy, Knowl.-Based Syst. 91:71–83.
[17] Pierpaolo D'Urso (2017) Informational Paradigm, management of uncertainty and theoretical formalisms in the clustering framework:A review, Inf. Sci. 400: 30-62.
[18] Peters G, Crespo F, Lingras P (2013) Weber R, Soft clustering-fuzzy and rough approaches and their extensions and derivatives. Int. J. Approx. Reasoning. 54 :307–22.

[19] Yu H (2018) Three-way decisions and three-way clustering, in: Rough Sets: International Joint Conference (IJCRS), Springer , 13-28.
[20] Wang PX, Yang XB, Ding WP, et al. (2024) Three-way clustering: Foundations, survey and challenges, Appl. Soft Comput. 151: 111131.
[21] Yao YY (2011) The superiority of three-way decisions in probabilistic rough set models, Information Sciences 181 :1080-1096.
[22] Yao YY (2018) Three-way decision and granular computing, Int. J. Approx. Reasoning. 103:107-123.
[23] Yao YY(2010) Three-way decisions with probabilistic rough sets, Inf. Sci. 180:341-353.
[24] Zhao TN, Zhang YJ, Miao DQ, Pedrycz W (2022) Selective label enhancement for multi-label classification based on three-way decisions, Int. J. Approx. Reasoning. 150:172-187.
[25] Wei L, Liu L, Qi JJ, Qian T (2020) Rules acquisition of formal decision contexts based on three-way concept lattices, Inf. Sci. 516:529-544.
[26] Zhang XY, Yao YY (2022) Tri-level attribute reduction in rough set theory, Expert Syst. Appl. 190:116187.
[27] Chen YX, Zhu P (2022) Three-way recommendation for a node and a community on social networks, Int. J. Mach. Learn. Cybern. 13:2909-2927.
[28] Yao JT, Azam N (2015) Web-based medical decision support systems for three-way medical decision making with game-theoretic rough sets, IEEE Trans. Fuzzy Syst. 23: 3-15.
[29] Yu B, Xie HJ, Fu Y, et al.(2024) Three-way graph convolutional network for multi-label classification in multi-label information system, Appl. Soft Comput. 161: 111767.
[30] Lang GM, Ding WP, Miao DQ, et al. (2024) Trisection-fusion and fusion-trisection methods of three-way conflict analysis with Pythagorean fuzzy information, Appl. Soft Comput. 164: 111939.
[31] Ren RS, Qi JJ, Wei, L, et al. (2024) Tri-level conflict analysis from the angle of three-valued concept analysis, Inf. Sci. 662: 120284.
[32] YU H, Chen LY, Yao JT (2021), A three-way density peak clustering method based on evidence theory. Knowl.-Based Syst. 211:106532.
[33] Chen YX, Zhu P, Yao YY (2024) An axiomatic framework for three-way clustering, Inf. Sci. 22:120761.
[34] Ali B, Azam N, Shah A, Yao JT (2021) A spatial filtering inspired three-way clustering approach with application to outlier detection, Int. J. Approx. Reason. 130:1-21.
[35] Wang PX,Shi H,Yang YB, et al. (2019) Three-way k-means: integrating k-means and three-way decision, Int. J. Mach. Learn. Cybern. 10:2767-2777.
[36] Chu XL, Sun BZ, Li X, et al. (2020) Neighborhood rough set-based three-way clustering considering attribute correlations: An approach to classification of potential gout groups, Inf. Sic. 535:28-41.
[37] Jiang CM, Li ZC, Yao JT (2022) A shadowed set-based three-way clustering ensemble approach, Int. J. Mach. Learn. Cyber. 13:2545-2558.
[38] Afridi MK, Azam N, Yao JT, Alanazi E (2018) A three-way clustering approach for handling missing data using GTRS, Int. J. Approx. Reason. 98:11-24.
[39] Yu H, Chang ZH, Wang GY, et al. (2020) An efficient three-way clustering algorithm based on gravitational search, Int. J. Mach. Learn. Cyber.11: 1003-1016.
[40] Jia XY, Rao Y, Li WW, et al.(2021) An automatic three-way clustering method based on sample similarity, Int. J. Mach. Learn. Cybern. 12:1545-1556.
[41] Zhang RT, Ma XL, Zhan JM, Yao YY, (2023) 3WC-D: A feature distribution-based adaptive three-way clustering method. Appl. Intell. 53:15561-15579.
[42] Wang PX, Wu TF, Yao YY (2023) A three-way adaptive density peak clustering (3W-ADPC) method, Appl. Intell. 53:23966-23982.
[43] Wang PX, Yao YY (2018) CE3: A three-way clustering method based on mathematical morphology, Knowl.-Based Syst. 155:54-65.
[44] Shah A, Azam N, Alanazi E, et al. (2022) Image blurring and sharpening inspired three-way clustering approach, Appl. Intell. 52:18131-18155.
[45] Yu H, Wang XC, Wang GY, et al. (2020) An active three-way clustering method via low-rank matrices for multi-view data, Inf. Sci. 507:823-839.
[46] Du MJ, Zhao JQ, Sun JR, Dong YQ (2023) M3W: Multistep three-way clustering, IEEE Trans. Neural Netw. Learn. Syst. https://doi.org/10.1109/TNNLS.2022.3208418.
[47] Wang PX, Yang XB (2021) Three-way clustering method based on stability theory, IEEE Access 9:33944-33953.
[48] Guo QH, Yin ZY, Wang PX (2022) An improved three-way k-means algorithm by optimizing cluster centers, Symmetry Basel 14: 1821.
[49] Khan S, Khan O, Azam N, et et al. (2023) Improved spectral clustering using three-way decisions, Inf. Sci. 641:119113
[50] Fan JS, Wang PX, Jiang C, Yang XB, Song JJ (2022) Ensemble learning using three-way density-sensitive spectral clustering, Int. J. Approx. Reason. 149:70-84.

[51] Sun C, Du MJ, Sun JR, et al. (2023) A three-way clustering method based on improved density peaks algorithm and boundary detection graph, Int. J. Approx. Reason. 153:239-257.
[52] Xu DL, Wang Y (2023) Density estimation for toroidal data using semiparametric mixtures, Stat. Comput. 33:140. doi:10.1007/s11222-023-10305-4.
[53] Duong T (2007) KS: Kernel density estimation and kernel discriminant analysis for multivariate data in R, J. Stat. Softw. 21:1-16.
[54] Li JZ, Yang XB, Song XN, et al. (2019) Neighborhood attribute reduction: a multi-criterion approach, Int. J. Mach. Learn. Cybern. 10: 731-742.
[55] Bouguettaya A, Yu Q, Liu XM, et al. (2015) Efficient agglomerative hierarchical clustering, Expert Syst. Appl. 42:2785-2797.
[56] Li Y, Zhou WJ (2022) A novel fuzzy distance-based minimum spanning tree clustering algorithm for face detection, Cogn. Comput. 14:1350-1361.
[57] Wang XC, Wang XL, Wilkes M (2009) A divide-and-conquer approach for minimum spanning tree-based clustering, IEEE Trans. Knowl. Data En. 21:945-958.
[58] Moshe Lichman, et al. (2013). UCI Machine Learning Repository. Irvine, CA.
[59] Liu Y H , Ma Z M, Yu F (2017). Adaptive density peak clustering based on K-nearest neighbors with aggregating strategy, Knowl.-Based Syst. 133: 208-220.