symmetry

MDPI

Article

An Improved Three-Way K-Means Algorithm by Optimizing Cluster Centers

Qihang Guo [1], Zhenyu Yin [1] and Pingxin Wang [2,*]

[1] School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212003, China
[2] School of Science, Jiangsu University of Science and Technology, Zhenjiang 212003, China
* Correspondence: wangpingxin@just.edu.cn

Abstract: Most of data set can be represented in an asymmetric matrix. How to mine the uncertain information from the matrix is the primary task of data processing. As a typical unsupervised learning method, three-way k-means clustering algorithm uses core region and fringe region to represent clusters, which can effectively deal with the problem of inaccurate decision-making caused by inaccurate information or insufficient data. However, same with k-means algorithm, three-way k-means also has the problems that the clustering results are dependent on the random selection of clustering centers and easy to fall into the problem of local optimization. In order to solve this problem, this paper presents an improved three-way k-means algorithm by integrating ant colony algorithm and three-way k-means. Through using the random probability selection strategy and the positive and negative feedback mechanism of pheromone in ant colony algorithm, the sensitivity of the three k-means clustering algorithms to the initial clustering center is optimized through continuous updating iterations, so as to avoid the clustering results easily falling into local optimization. Dynamically adjust the weights of the core domain and the boundary domain to avoid the influence of artificially set parameters on the clustering results. The experiments on UCI data sets show that the proposed algorithm can improve the performances of three-way k-means clustering results and is effective in revealing cluster structures.

Keywords: three-way decision; three-way clustering; three-way k-means; ant colony

MSC: 68T37

Citation: Guo, Q.; Yin, Z.; Wang, P. An Improved Three-Way K-Means Algorithm by Optimizing Cluster Centers. *Symmetry* **2022**, *14*, 1821. https://doi.org/10.3390/sym14091821

Academic Editor: José Carlos R. Alcantud

Received: 20 July 2022
Accepted: 24 August 2022
Published: 2 September 2022

1. Introduction

Rapid developments of science and technology produce a large number of data anytime and anywhere in modern society. How to mine valuable information from massive data has been a challenging task of information science and artificial intelligence. As one emerging technology of intelligence information processing, granular computing [1,2] deals with information in the form of some aggregates and their processing. The main task of granular computing is to construct different granular structures by various methods of information granulation [3]. There are many different approaches of information granulation, among which, clustering analysis is one of the widely and most used one [4]. Clustering analysis [5] is a multivariate analysis method in statistics. The objective of clustering is to divide a set of objects into different groups such that the objects in the same cluster have high similarity while the objects in the different groups have high dissimilarity. As a powerful data analysis technique, clustering has been widely used information granulation [6–8], information fusion [9–11], attribute reduction [12–15], feature selection [16–18], and other fields.

There are many different clustering algorithms, among which, k-means [19] is one of the most commonly used one. The object of k-means algorithm is to minimize the following function,

$$J = min \sum_{j=1}^{k} \sum_{v_i \in C_j} ||v_i - \mu_j||^2, \quad (1)$$

where $||v_i - \mu_j||$ is a chosen distance measure between a data point v_i and the cluster centre μ_j of j-th cluster. The process of k-means has three steps. The first one is to randomly select k samples as the initial cluster centers, the second one is to calculate the distance between each sample and cluster center, and the third one is to assign the samples to the nearest cluster. The process is repeated by continuously updating the iterative cluster centers until a certain termination condition is met. Although k-means algorithm has been widely used since it was proposed, it still has the following problems.

1. The clustering results of k-means are dependent on the random selection of clustering centers and the problem of local optimization readily occurs.
2. Traditional k-means algorithms are based on the assumption that a cluster is represented by one single set with a sharp boundary. Only two types of relationship between an object and cluster are considered, i.e., belong to and not belong to. The requirement for a sharp boundary is easily met for analytical clustering results, but may not adequately show the uncertainty information in the dataset.

To solve the above problems, many methods have been developed to improve the results of the k-means algorithm. For example, Zhang et al. [20] proposed an improved k-means algorithm based on a density canopy to solve the problem of determining the best initial seeds. Wang et al. [21] presented a three-way k-means algorithm by integrating three-way decisions [22,23] into clustering to depict the uncertainty information in the dataset. The main idea of the three-way k-means algorithm is to introduce fault-tolerant errors in the k-means iteration process, and represents the results of each cluster with a core region and a fringe region. However, in common with the k-means algorithm, the three-way k-means algorithm also has the problems that the clustering results are dependent on the random selection of clustering centers, are sensitive to noise points and outliers, and easily succumb to local optimization.

To solve the local optimal problem of the three-way k-means algorithm, this paper presents an improved three-way k-means algorithm by integrating the ant colony [24] algorithm into the k-means algorithm. The ant colony algorithm simulates the foraging behavior of ants in nature, and initially solved a problem of traveling salesmen by analogy with a pheromone mechanism. Its inspiration came from the behavior of ants in finding the shortest path in the process of looking for food. Compared with other clustering algorithms, the ant colony algorithm has the advantages of strong robustness, good parallelism, adaptability to computer systems, ease of combination with other algorithms and others.

The positive pheromone feedback mechanism in the ant colony algorithm can be transformed into a clustering problem. In the ant colony clustering algorithm, based on ant colony foraging behavior, food sources are regarded as different clustering centers, and data are regarded as ants with different attributes. Under the guidance of pheromones, ants move between different food sources with a certain probability, and finally form clustering results around different food sources. In the proposed algorithm, we use the random probability selection strategy of ant decision-making in the ant colony algorithm to optimize the initial clustering centers. On this basis, iterative updating can better solve the problem of the sensitivity of the three-way k-means algorithm to the initial clustering center.

This article is structured in the following manner: In Section 2, we introduce the background to our proposed methods, including three-way clustering and three-way k-means. The detailed process of the proposed algorithm is presented in Section 3. In Section 4, experiments are described which evaluate the performance of the proposed algorithms. Some concluding remarks and discussion of future research directions are presented in Section 5.

2. Related Work

2.1. Three-Way Clustering

In classical two-way decision clustering, a cluster is represented by a set with a crisp boundary. There are only two relationships between an object and a class: the object either belongs or does not belong to the class. The requirement of a crisp boundary is conducive to analysis results, but may not adequately adequately show the uncertainty information in the data set. In order to address the problem of information uncertainty, Yao [22,23] proposed a theory of three-way decision by extending the commonly used binary-decision model. There are three main tasks in three-way decision [25]: (1) trisecting strategies, (2) acting strategies, and (3) outcome evaluation. The whole process can be depicted as TAO (trisecting-acting-outcome) [25] framework shown in Figure 1.

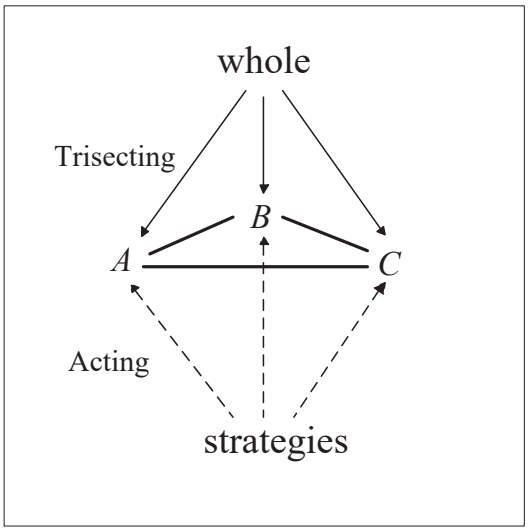


Figure 1. The TAO of three-way decision (adapted from [25]).

With the development of three-way decision, many applications of three-way decision were researched in various fields, such as, data analysis [26,27], concept analysis [28–30], granular computing [31,32], sequential three-way decision [33,34].

To relax this requirement of sharp boundary, a new type of clustering algorithms [35,36] was proposed, named three-way clustering. Unlike traditional hard clustering, three-way clustering presents a cluster through a pair of sets:

$$C_i = (Co(C_i), Fr(C_i)), \tag{2}$$

where $Co(C_i) \subset V$ and $Fr(C_i) \subset V$ are defined as the core region and fringe region of cluster C_i, respectively. These two sets divide the universe into three parts $Co(C_i)$, $Fr(C_i)$ and $Tr(C_i)$, which chapter the three types of relationships between the objects and cluster, namely, objects belong to the cluster, objects not belong to the cluster, and objects partially belong to the cluster. For the samples in $Co(C_i)$, they belong to the cluster (C_i) definitely and have a higher within-class similarities. For the samples in $Fr(C_i)$, they maybe belong to the cluster (C_i) and have a lower similarities with the core samples. For the samples are in $Tr(C_i)$, they do not belong to the cluster C_i definitely. Different from using one single set to represent the cluster, three-way clustering uses $Co(C_i), Fr(C_i)$ to represent one cluster. Three-way clustering addresses the problem of information uncertainty in traditional clustering methods through adding fringe region, which reduces decision risk caused by inaccurate information. We take Figure 2 as an example to show one three-way clustering result.

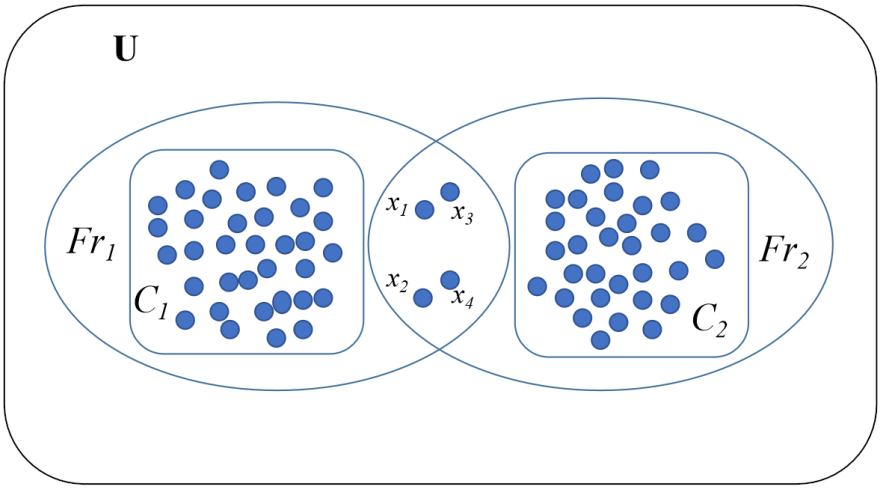


Figure 2. An illustrative three-way clustering result.

There are many different strategies to obtain the the core regions and fringe regions of three-way clustering. Typically, evaluation function is a commonly used method. The main idea of evaluation function is to construct a function and a pair of thresholds. The function assigns each sample a value. The samples with values greater than one threshold are assigned to the core region and the objects with values between the two thresholds are assigned to the fringe region. From a statistical point of view, a sample belongs to a certain cluster with a certain probability. It is reasonable to assign a sample to core region of one cluster when the probability of the sample belonging to this cluster is much greater than the probability of the sample belonging to other clusters. However, when the probability of a sample belonging to some clusters are almost same, it will be difficult to make a positive decision. Adopting a delayed decision and assigning them into the fringe region can reduce the risk of decision. Three-way clustering addresses the problem of information uncertainty in traditional clustering methods through adding fringe region, which reduces decision risk caused by inaccurate information. Recently, three-way clustering has attracted a lot of research, and many three-way clustering algorithms were developed. Wang et al. [37] proposed a three-way clustering framework by using contraction and expansion operators inspired in mathematical morphology; Jia et al. [38] introduced the definition of sample similarity to measure the uncertainty and developed an automatic three-way clustering approach. Fan et al. [39] proposed three-way density-sensitive spectral clustering algorithm by using density-sensitive metric. Shah [40] proposed a new three-way clustering by using image inspired cluster blur and sharp operators. Except the above research results, some other algorithms also enrich the theories and models of three-way clustering [41–45].

Given a set of data objects $U = \{x_1, x_2, \cdots, x_n\}$, $C = \{c_1, c_2, \cdots, c_k\}$ is a finite set of clusters, and U is divided into k classes. The idea of three-way clustering is to use a pair of sets to represent each cluster. This pair of sets consists of a core region (Co), a fringe region (Fr), and a trivial fringe (Tr) [46]. The results of three-way clustering can be represented as the following family of clusters:

$$\mathbb{C} = \{(Co(C_1), Fr(C_1)), (Co(C_2), Fr(C_2)), \cdots, (Co(C_k), Fr(C_k))\}. \tag{3}$$

According to the definition of clustering results, $Co(C_i)$ and $Fr(C_i)$ must meet the following three conditions:

$$(1) Co(C_i) \neq \phi, i = 1, 2, \cdots, k; \tag{4}$$

$$(2) \bigcup_{i=1}^{k} (Co(C_i) \cup Fr(C_i)) = U; \tag{5}$$

$$(3) Co(C_i) \cap Co(C_j) = \phi, i \neq j. \tag{6}$$

2.2. Three-Way k-Means

The traditional k-means algorithm is an iterative clustering analysis algorithm. The first step of the k-means algorithm is to randomly select k samples as the initial cluster centers. The second step is to calculate the distance between each sample and cluster center, and then assign the samples to the nearest cluster. The process is repeated by continuously updating the iterative cluster centers until a certain termination condition is met. In the process of k-means algorithm, there are only two relationships between samples and various clusters: belonging to the cluster and not belonging to the cluster. This two relationships ignore the samples which partially belong to one cluster. In fact, there is another type of relationship between sample and various clusters: belong to partially. In order to capture these three types of relationships, three-way k-means [21] was proposed by integrating three-way decision theory with the k-means algorithm. The main idea of the three-way k-mean algorithm is to introduce fault-tolerant errors in the k-means iteration process, and to represent the result of each cluster with a core region and a fringe region. The procedure of three-way k-means clustering consists mainly of two steps. The first step is to obtain the support of each cluster and the second step is to separate the core region from the support. For each object v and randomly selected k centroids $x_1, \cdots, x_k$, let $d(v, x_j)$ be the distance between itself and the centroid x_j. Suppose $d(v, x_i) = \min_{1 \le j \le k} d(v, x_j)$ and $T = \{j : d(v, x_j) - d(v, x_i) \le \varepsilon_1 \text{ and } i \ne j\}$, where ε_1 is a given parameter. Then,

1. If $T \ne \phi$, then $v \in \text{support}(C_i)$ and $v \in \text{support}(C_j)$.
2. If $T = \phi$, then $v \in \text{support}(C_i)$.

The modified centroid calculations for the above procedure are given by:

$$x_i = \frac{\sum_{v \in \text{support}(C_i)} v}{|\text{support}(C_i)|}, \tag{7}$$

where $i = 1, \cdots, k$, v are all objects in $\text{support}(C_i)$, and $|\text{support}(C_i)|$ is the number of objects in $\text{support}(C_i)$.

This process is repeated until modified centroids in the current iteration are identical to those that have been generated in the previous one, i.e., when the prototypes are stabilized. The second step is to separate the core regions from the supports using a perturbation analysis method. Algorithm 1 is designed to describe the process of TWKM clustering.

Algorithm 1: Three-way k-means [21]

Input: A set of points $V = \{v_1, \cdots, v_n\}$, the number of clusters k and parameters ε_1 and ε_2.

Output: $\mathbb{C} = \{(\mathrm{Co}(C_1), \mathrm{Fr}(C_1)), (\mathrm{Co}(C_2), \mathrm{Fr}(C_2)), \cdots, (\mathrm{Co}(C_k), \mathrm{Fr}(C_k))\}$.

randomly select k cluster centroids $x_1, \cdots, x_k$
For $i = 1 : n$ **do**
  for a data point v_i, determine its closest centroid $x_h : d(v_i, x_h) = \min_{1 \le c \le k} d(v_i, x_c)$;;
  determine set $T = \{j : d(v_i, x_j) - d(v_i, x_h) \le \varepsilon_1 \text{ and } j \ne h\}$;
  If $T = \phi$
    assign x_i to the support of the cluster h, i.e., $v_i \in \mathrm{support}(C_h)$;
  Else
    assign x_i to the support of the cluster h and j, i.e., $v_i \in \mathrm{support}(C_h)$ and $v \in \mathrm{support}(C_j)$;
  End
  calculate the new centroid for each cluster using Equation (5);
  until modified centroids in the current iteration are identical to those of the previous.
End
For $i \leftarrow 1$ *to* k **do**
  for each $v \in \mathrm{support}(C_i)$, determine set $H = \{j : j \ne i \wedge v \in \mathrm{support}(C_j)\}$;
  If $H \ne \phi$
    assign v to the fringe region of C_i, i.e., $v \in \mathrm{Fr(C_i)}$;
  Else
    add m_i times v into $\mathrm{support}(C_i)$ and denote the new cluster by $\mathrm{support}(C_i^*)$, where m_i is the number of elements in $\mathrm{support}(C_i)$;
    calculate the new centroid x_i^* of $\mathrm{support}(C_i^*)$ by Equation (7) and the differences $|x_i^* - x_i|$; **If** $|x_i^* - x_i| \le \varepsilon_2$
      assign v to the core region of C_i, i.e., $v \in \mathrm{Co(C_i)}$;
    Else
      assign v to the fringe region of C_i, i.e., $v \in \mathrm{Fr(C_i)}$;
    End
  End
End
Return $\{(\mathrm{Co}(C^1), \mathrm{Fr}(C^1)), (\mathrm{Co}(C^2), \mathrm{Fr}(C^2)), \cdots, (\mathrm{Co}(C^K), \mathrm{Fr}(C^K))\}$.

3. The Improved Three-Way k-Means

The three-way k-means clustering algorithm integrates three-way decision theory with the k-means algorithm and uses a pair of sets to represent a cluster, which can effectively deal with the uncertainty of data. However, as for the k-means algorithm, the three-way k-means method is still sensitive to the initial clustering centers and can easily succumb to the problem of local optimization.To solve this problem, we present an improved three-way k-means clustering algorithm by combining a random probability selection strategy and the pheromone feedback mechanism in the ant colony algorithm with three-way k-means. The sensitivity of the three k-means clustering algorithm to the initial clustering center is optimized through continuous updating iterations, so as to avoid the clustering results easily falling into local optimization. The weights of the core domain and the boundary domain are dynamically adjusted to avoid the influence of artificially set parameters on the clustering results.

3.1. Random Probability Selection Strategy

The ant colony algorithm simulates the foraging behavior of ants in nature, and initially solves the problem of traveling salesman by analogy with a pheromone mechanism.

The inspiration for the ant colony algorithm came from the behavior of ants in finding the shortest path in the process of looking for food. In the ant colony algorithm, the process of ants looking for food sources can be viewed as continuous clustering. According to the size of the current pheromone quantity, ants randomly select according to probability, and allocate samples to each cluster center. The larger the pheromone amount between a sample and a cluster center, the greater the probability that the sample will be assigned to this class. In the process of clustering, the probability that a sample is assigned to a cluster center is calculated according to the size of the pheromone between the sample and the cluster center and the heuristic function. The random probability selection strategy greatly increases the effectiveness of the algorithm, causes the algorithm to have the characteristics of convergence, and prevents the algorithm from falling into local optimization. The probability calculation formula of the cluster center to which the ant search sample belongs is:

$$p_{ij} = \tau_{ij}^{\alpha}(t)\eta_{ij}^{\beta}(t), \tag{8}$$

where $\eta_{ij}(t)$ is a heuristic function and $\eta_{ij} = \frac{1}{d_{ij}}(i = 1, 2, \ldots n; j = 1, 2, \ldots, k)$, d_{ij} represents sample v_i to cluster center x_j. $\tau_{ij}(t)(i = 1, 2, \ldots n; j = 1, 2, \ldots, k)$ represents the pheromone concentration between sample v_i and cluster center x_j. The pheromone concentration is distributed between the sample and the clusters, and the initial pheromone concentration is 1, t is the number of iterations, α is the pheromone importance factor, and β is the heuristic importance factor. To increase the diversity of search and to speed up the convergence speed, at the beginning, ants randomly select a sample v_i as the starting point, and then Formula (7) is used to calculate the probability p of the sample to each cluster center x_j. The cluster of sample v_i is determined using roulette. The above process is repeated for another sample until all the samples are traversed to form a solution.

In the ant colony algorithm, the objective function is used to evaluate the solution formed by all ants after completing an iteration, and only the clustering results obtained by the ants with the best objective function value are retained. We construct the fitness function using the intra-cluster cohesion function and the inter-cluster dispersion function, so that the objects in the same cluster are as similar as possible, and the objects in different clusters are as different as possible. The intra-cluster cohesion function is defined as follows,

$$J = \sum_{i=1}^{k}(\omega_{icore} \sum_{x_i \in icore} (\|v_i - m_k\|^2) + (\omega_{ifringe} \sum_{x_i \in ifringe} (\|v_i - m_k\|^2), \tag{9}$$

where J represents the sum of the distance between each sample v_i and cluster center x_j, which is used to evaluate the degree of cohesion. ω_{icore} and $\omega_{ifringe}$ represent the weight values of the core region and the fringe region, respectively. Dynamic adjustment ω_{icore} and $\omega_{ifringe}$ can effectively avoid the influence of a sample's number change in the core region and the fringe region, and can also avoid the influence of clustering centers due to differences in distance distribution. In this paper, we assume that ω_{icore} and $\omega_{ifringe}$ satisfy the following equations.

$$\omega_{icore} + \omega_{ifringe} = 1, \tag{10}$$

$$\frac{\omega_{icore}}{\omega_{ifringe}} = \frac{|\omega_{icore}|}{|\omega_{icore}|}, \tag{11}$$

ω_{icore} and $\omega_{ifringe}$ where $|\omega_{icore}|$ and $|\omega_{icore}|$ are the number of samples in the core region and the fringe region, respectively.

The quality of clustering results is determined by the intra-cluster distance and the inter-cluster distance; when the intra-class distance is smaller, the inter-class distance is larger, the value of the objective function is smaller, and the clustering result is better. The inter-cluster dispersion function is defined as,

$$D = \frac{1}{k}\left(\sum_{i,j=1}^{k} \|x_i - x_j\|^2\right), \tag{12}$$

where x_i and x_j represent the cluster center of C_i and C_j, respectively.

Based on the intra-cluster cohesion function and the inter-cluster dispersion function defined by (8) and (10), we construct the following fitness function to optimize three-way k-means:

$$Fit = \frac{1+D}{1+J}. \tag{13}$$

In the process of looking for food sources, ants release pheromones on the paths they pass. The higher the pheromone concentration, the shorter the distance of the road. In this way, the more ants walk, the higher the pheromone concentration on this path. Each ant moves towards the direction with the highest pheromone concentration, and the continuously strengthened pheromone attracts more ants, so a positive feedback mechanism is formed. As time goes on, the pheromone on the poor path cannot be strengthened, and as the pheromone volatilizes continuously, it loses its attraction, thus forming a negative feedback mechanism.

The positive feedback mechanism attracts more ants to choose the current optimal path, accumulates more pheromone, increases the probability of other ants choosing the path, narrows the scope of ant search, and promotes the convergence of the clustering algorithm. The negative feedback mechanism can eliminate the effect of the positive feedback mechanism, effectively preventing more ants from being attracted to the optimal path, and making the algorithm result fall into the local optimal solution. Using the positive and negative feedback of pheromones, the ant colony algorithm avoids the algorithm falling into a local optimal solution and increases the diversity of solutions.

Pheromone updating in the ant colony algorithm uses the overall information of the ant colony. When the ant releases the pheromone, the pheromone remaining on the path will gradually disappear. This is also to make the next generation of ants more robust both globally and locally when choosing the path. Therefore, when all the ants have completed a cycle, the global update of the residual pheromone is carried out. The pheromone update formula in the ant colony algorithm is as follows:

$$\tau_{ij}(t+1) = (1-\rho) * \tau_{ij}(t) + \Delta\tau_{ij}, \tag{14}$$

$$\Delta\tau_{ij} = 0.1 * \tau_{ij}, \tag{15}$$

where, parameter $0 < \rho < 1$ indicates the degree of volatilization of the pheromone, τ_{ij} represents the pheromone concentration, and $\Delta\tau_{ij}$ represents the increment in pheromone.

3.2. The Improved Three-Way k-Means Algorithm

Because the clustering results of the standard three-way k-means algorithm depend on the selection of initial centers they easily succumb to the problem of local optimization. To overcome this problem, this paper presents an improved three-way k-means algorithm by integrating the ant colony algorithm and three-way k-means. An original element of this paper is the application of clustering centers obtained by the ant colony algorithm to the three-way k-means, which makes up for the shortcomings of the three-way k-means clustering algorithm due to the random selection of clustering centers. Figure 3 presents a flowchart of the proposed algorithm.

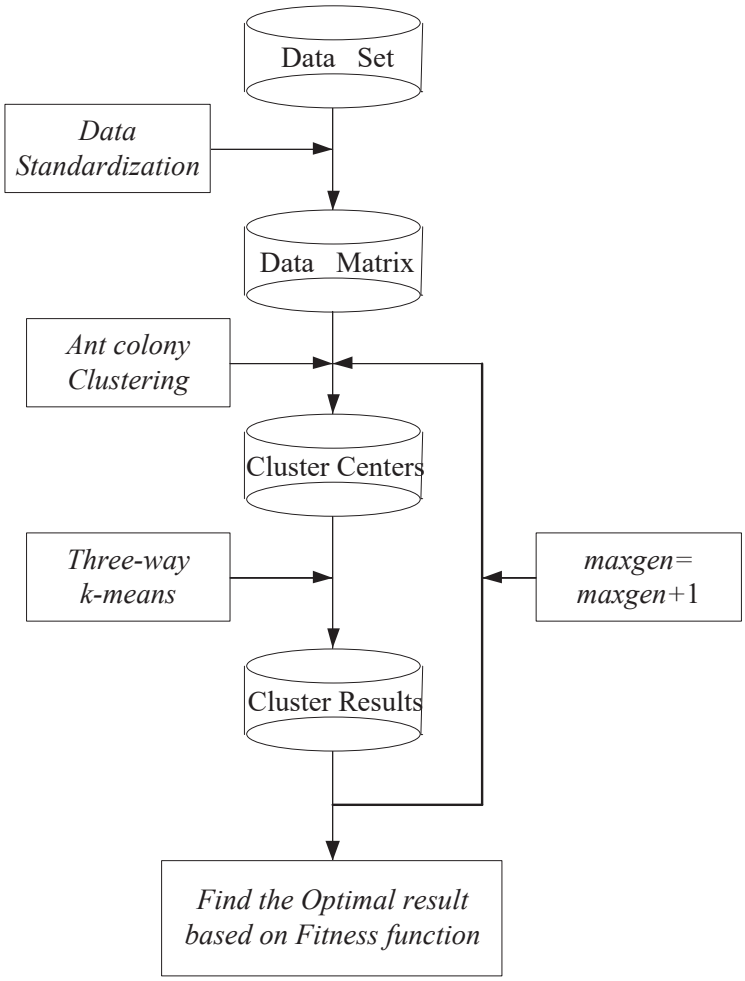


Figure 3. A flowchart of the proposed algorithm.

The ant colony algorithm simulates the foraging behavior of ants in nature, and initially solves the problem of traveling salesman by analogy with a pheromone mechanism. Ants randomly select a sample in the sample space as the starting point. The probability that a sample is assigned to a cluster is obtained according to the amount of the pheromone between the sample and the cluster center. The sample is allocated to a cluster by roulette. Then the ant selects another sample until all samples are assigned, that is, when an iteration is completed to form a solution. The optimal solutions are calculated using the value of the objective function. The pheromone in the ant colony algorithm reflects the overall information in the ant colony. When the ant releases the pheromone, the pheromone remaining on the path will gradually disappear. This is also to make the next generation of ants more robust globally and locally when choosing the path. The specific steps of the algorithm are shown in Algorithm 2:

The detailed complexity of Algorithm 2 is as following: Line 3 to Line 11 are to find the support of each cluster. The time complexity of this process is $O(knm)$, where n and m are the number of elements and attributes, respectively. Line 12 is to separate the core regions from the support sets using centroid perturbation analysis. The time complexity of this process is $O(knm)$. Line 13 to Line 14 are to update the process. The time complexity of Algorithm 2 is $O(tknm) + O(knm)$, where t is the number of iterations.

Algorithm 2: The improved three-way k-means based on ant colony algorithm.

Input: A set of points $V = \{v_1, \cdots, v_n\}$, the number of clusters k, parameters $maxgen$ and q.
Output: $\mathbb{C} = \{(\mathrm{Co}(C_1), \mathrm{Fr}(C_1)), (\mathrm{Co}(C_2), \mathrm{Fr}(C_2)), \cdots, (\mathrm{Co}(C_k), \mathrm{Fr}(C_k))\}$.

randomly select k cluster centroids $x_1, \cdots, x_k$;
For $j = 1 : maxgen$ **do**
  For $i = 1 : n$ **do**
    calculate the probability of the sample v_i selected by each ant to each cluster center to obtain the set $p = p_1, p_2, ..., p_k$ by Equation (8);
    select the maximum probability p_{max} for v_i and assign v_i to the upper bound of corresponding cluster C^u_{max} ;
    calculate the differences p_{poor} between the p_{max} and the rest points in the set p;
    If $p_{poor} < q$
      assign v_i to the upper bound of corresponding cluster too;
    End
    calculate the new centroid for each cluster using Equation (7);
  End
  obtain the core region and the fringe region of each cluster by steps 13–25 of Algorithm 1;
  calculate the value of the fitness function by Equations (9), (12) and (13);
  update pheromone by Equations (14) and (15);
End
If *the algorithm reaches the maximum number of iterations maxgen*
  return the optimal solution
Else
  $maxgen = maxgen + 1$
End
Return $\{(\mathrm{Co}(C_1), \mathrm{Fr}(C_1)), (\mathrm{Co}(C_2), \mathrm{Fr}(C_2)), \cdots, (\mathrm{Co}(C_k), \mathrm{Fr}(C_k))\}$.

4. Experimental Analysis

4.1. Evaluation Indices

1. Accuracy (Acc) [47]

$$Acc = \frac{1}{n} \sum_{i=1}^{k} C_i \tag{16}$$

In the above formula, n is the total number of samples in the dataset, C_i is the correct number of samples divided into class clusters i, and k is the number of class clusters. Acc represents the ratio between the number of correctly partitioned elements and the total number. A greater ACC value implies a better clustering result. When ACC = 1, the result of the clustering algorithm is consistent with the real result.

2. Davies–Bouldin index (DBI) [47].

$$DB = \frac{1}{c} \sum_{i=1}^{c} \max_{j \neq i} \left\{ \frac{S(C_i) + S(C_j)}{d(x_i, x_j)} \right\} \tag{17}$$

where $S(C_i)$ and $d(x_i, x_j)$ are the intra-cluster distance and the inter-cluster separation, respectively. $S(C_i)$ is defined as follows:

$$S(C_i) = \frac{\sum_{v \in C_i} \| v - x_i \|}{| C_i |}. \tag{18}$$

As a function of the ratio of the within cluster scatter to the between cluster separation, a lower value will mean that the clustering is better.

3. Average silhouette index (AS) [47].

$$AS = \frac{1}{n}\sum_{i=1}^{n} S_i, \tag{19}$$

where n is the total number of objects in the set and S_i is the silhouette of object v_i, which is defined as,

$$S_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}, \tag{20}$$

a_i is the average distance between x_i and all other objects in its own cluster, and b_i is the minimum of the average distance between x_i and objects in other clusters. The range of the average silhouette index is $[-1, 1]$; a larger value means a better clustering result.

4.2. Performances of Proposed Algorithm

To test the performances of our proposed algorithm, we employed 10 datasets from the UCI Machine Learning repository, which were Wine, Class, Ecoli_Nor, Forest, Bank, Iris, Contraceptive, Molecular Biology1, Libras_Nor, and Caffeine_consumption. Table 1 shows the details of these datasets.

Table 1. A description of dataset used.

ID	Datasets	Samples	Attributes	Classes
1	Wine	178	13	3
2	Class	214	9	6
3	Ecoli	366	7	8
4	Forest	523	27	4
5	Bank	1372	4	2
6	Iris	150	4	3
7	Contraceptive	1473	9	3
8	Molecular Biology	106	52	2
9	Libras	360	90	15
10	Caffeine Consumption	1885	12	7

Because the evaluation indices Acc, DBI and AS are only adopted for the hard clustering results, three-way clustering cannot calculate these values directly. To present the performances of our proposed algorithm, we use the core regions to form a clustering result, then calculate the Acc, DBI and AS values using the core region to represent the corresponding cluster. The average ACC, DBI and AS values are achieved by running the process 30 times on all datasets. To compare clustering effects, the performances of k-means [19], FCM [48], and three-way k-means [21] are also presented in Tables 2–4. The best performance for each dataset is highlighted in bold.

Table 2. The performances of average *Acc* value.

ID	Data Sets	k-Means	FCM	Three-Way k-Means	Proposed Algorithm
1	Wine	0.6573	0.6692	0.6831	**0.6911**
2	Class	0.5981	0.6007	0.6112	**0.6366**
3	Ecoli	0.6339	0.6335	0.6652	**0.6773**
4	Forest	0.7795	0.7540	0.7807	**0.8294**
5	Bank	0.5758	0.5969	0.6123	**0.6131**
6	Iris	0.8866	0.8933	**0.9040**	**0.9040**
7	Contraceptive	0.2145	0.2179	0.2822	**0.2826**
8	Molecular Biology	0.6037	0.6226	0.6547	**0.6659**
9	Libras	0.8611	0.9162	**0.9256**	0.9240
10	Caffeine Consumption	0.2005	0.1960	0.2411	**0.2422**

Table 3. The performances of average *DBI* value.

ID	Data Sets	k-Means	FCM	Three-Way k-Means	Proposed Algorithm
1	Wine	1.7835	1.6922	1.5542	**1.5431**
2	Class	1.0475	1.2233	0.7855	**0.7596**
3	Ecoli	1.1504	1.0273	0.9667	**0.9425**
4	Forest	1.2774	1.2253	1.200	**1.1879**
5	Bank	1.1913	1.1952	1.1332	**1.1267**
6	Iris	0.7609	0.7507	0.7236	**0.7355**
7	Contraceptive	1.2716	1.2539	1.2323	**1.2220**
8	Molecular Biology	4.9588	4.8236	4.6783	**4.6689**
9	Libras	1.9240	1.9126	1.9033	**1.9023**
10	Caffeine Consumption	1.9116	1.8072	1.6655	**1.6048**

Table 4. The performances of average *AS* value.

ID	Data Sets	k-Means	FCM	Three-Way k-Means	Proposed Algorithm
1	Wine	0.3383	0.2337	0.3347	**0.3574**
2	Class	0.5309	0.5543	0.5887	**0.6038**
3	Ecoli	0.4419	0.4326	0.4433	**0.4524**
4	Forest	0.4029	0.4302	0.4559	**0.4669**
5	Bank	0.5000	0.4954	0.5111	**0.5280**
6	Iris	0.6959	0.7091	0.7114	**0.7188**
7	Contraceptive	0.4236	0.4309	0.4597	**0.4672**
8	Molecular Biology	0.0553	0.0538	0.0558	**0.0585**
9	Libras	0.3519	0.3000	0.3533	**0.3556**
10	Caffeine Consumption	0.3150	0.3491	0.3517	**0.3563**

To analyze the time comparison of different algorithms, Figure 4 lists the running time of different algorithms on the 10 UCI datasets where the unit of measurement for time is "second".

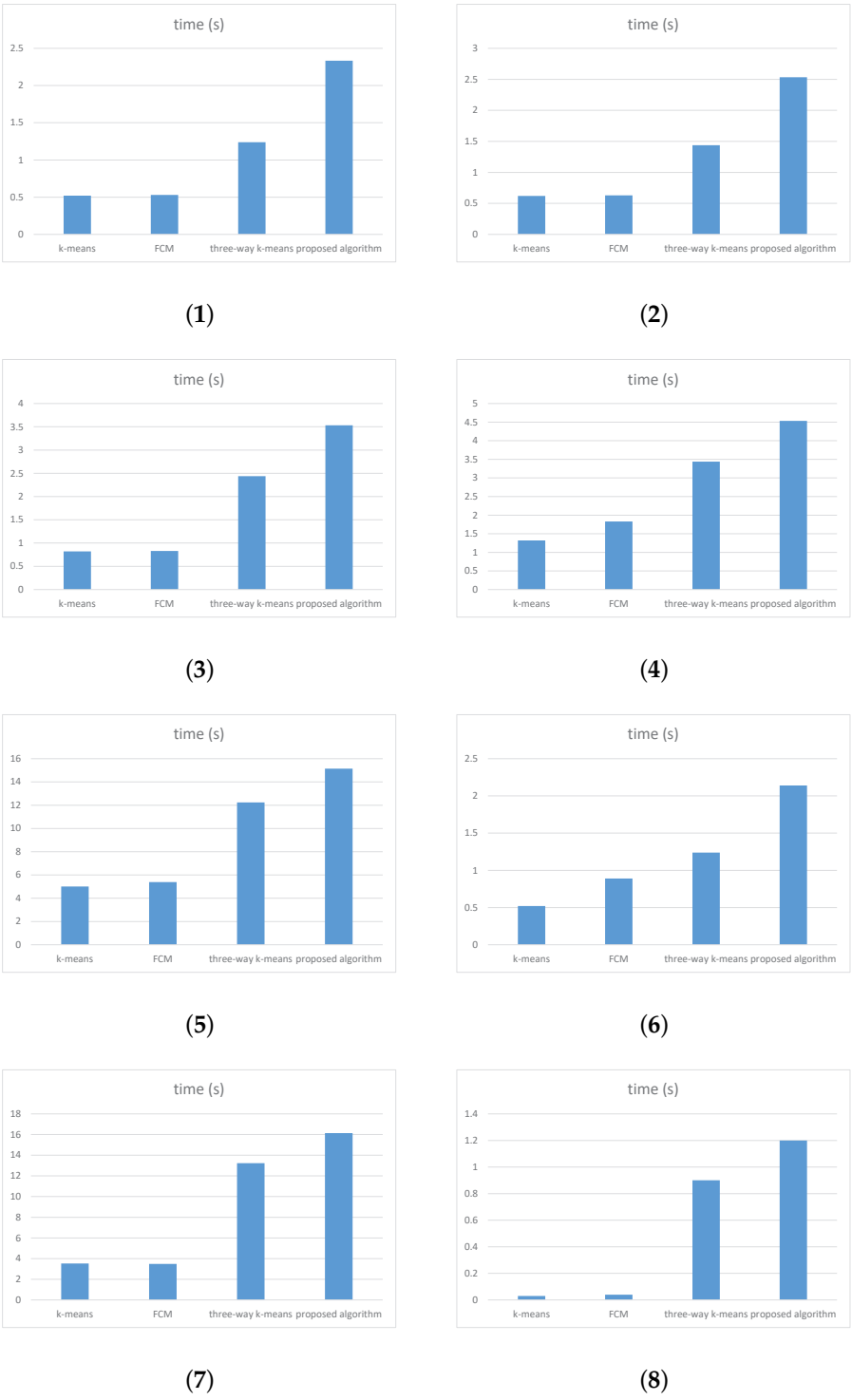
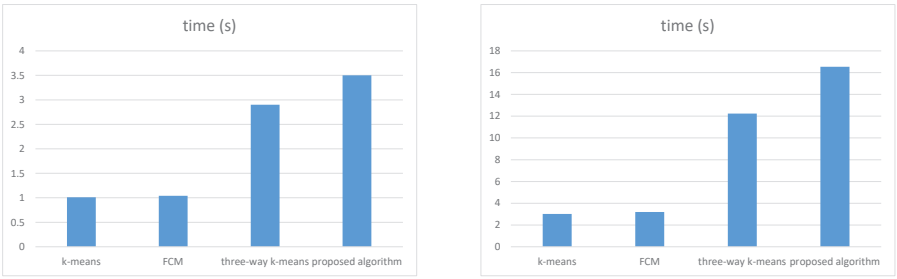


Figure 4. *Cont.*



(**9**) (**10**)

Figure 4. Time comparison of different algorithms on UCI datasets. (**1**) Wine. (**2**) Class. (**3**) Ecoli. (**4**) Forest. (**5**) Bank. (**6**) Iris. (**7**) Contraceptive. (**8**) Molecular Biology. (**9**) Libras. (**10**) Caffeine Consumption.

4.3. Experimental Results Analysis

The average Acc values of k-means, FCM, three-way k-means and the proposed algorithm are reported in Table 2. Obviously, comparing with other three algorithms, the proposed algorithm achieves a better ACC value on 8 data sets. Though the performances of Acc value on Iris and Libras by three-way k-means algorithm are equal to or superior to the results by the proposed algorithm, while the DBI and AS value of our proposed algorithm are better than the results of three-way k-means. The increase of Acc value indicates that the proposed algorithm assign more samples to right cluster than other algorithms.

The results of DBI and AS are listed in Tables 3 and 4, respectively. From Tables 3 and 4, we can find that the proposed algorithm obtains better results than other algorithms in terms of DBI and AS value on all data sets. Since DBI is the ratio of the within cluster scatter to the between cluster separation, the improvement of DBI means that the clustering results of proposed algorithm have a higher cluster separation. The results of AS can also verify the fact that the clustering results of proposed algorithm have a lower intra class distance and a higher inter class distance. This can be attributed to the fact that the clustering result of the proposed algorithm is represented by core regions when calculate the DBI value and AS value, which can increase the degree of separation between clusters and reduce the degree of dispersion within clusters.

The running time of the proposed algorithm is much longer than k-means and FCM on all the UCI datasets, as shown in Figure 4. This is because the time complexity of the proposed algorithm is much more than that of k-means and FCM. Compared with the three-way k-means method, the proposed algorithm takes slightly longer. This is because the proposed algorithm integrates the ant colony algorithm into three-way k-means, which adds to the running time of the proposed algorithm.

The above discussion suggests that the proposed algorithm can effectively improve the results of three-way k-means, solving the clustering problem of uncertain elements and maintaining improved clustering performance at the same time. Though the proposed algorithm can achieve better clustering results than three-way k-means, it still has the following two disadvantages:

1. Similar to the k-means algorithm, the proposed method can achieve good results for convex datasets. If the dataset is non-convex, the proposed algorithm fails to give good results.
2. The time complexity and computation complexity of the proposed algorithm are higher than for k-means and three-way k-means, which means it is not suitable for big data.

5. Conclusions and Future Work

Three-way clustering uses a core region and a fringe region to represent a cluster, which divide the universe into three disjoint sets to reflect the three types of relationship

between an object and a cluster. The samples in the core region are determined to belong to this type of cluster, while the samples in the fringe region may belong to this type of cluster. Three-way clustering assigns the samples with uncertainty information into corresponding fringe regions, which reduces the risk of decision-making. However, the standard three-way k-means algorithm does not always guarantee good results, as the accuracy of clustering depends on the selection of the initial centroids and easily succumbs to the problem of local optimization. Aiming to solving this problem, this paper presents an improved three-way k-means algorithm by integrating the ant colony algorithm and three-way k-means. The idea of this algorithm is to apply the clustering center obtained by each iteration of the ant colony algorithm to the three-way k-means, which compensates for the shortcomings of the three-way k-means clustering algorithm due to the random selection of clustering centers. The proposed algorithm optimizes three-way k-means using a random probability selection strategy in the ant colony algorithm and the positive and negative feedback mechanism of pheromones to dynamically adjust the weight. The experiments on UCI datasets show that the proposed algorithm can improve the performance of the three-way k-means clustering results according to the comparison of ACC, DBI and AS.

Finally, we should point out that the standard three-way k-means algorithm has two problems, the selection of the initial clustering centers and the determination of the cluster number. This paper presents a method to improve the selection of the initial clustering centers. The determination of cluster number is an interesting topic that requires further investigation.

Author Contributions: Conceptualization, P.W.; data curation, Q.G.; formal analysis, Z.Y.; funding acquisition, P.W.; methodology, P.W.; supervision, P.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of China (nos. 62076111, 62006099) and the Key Laboratory of Oceanographic Big Data Mining & Application of Zhejiang Province (no. OBDMA202002), Postgraduate Research & Practice Innovation Program of Jiangsu Province (nos. KYCX22_3826, SJCX22_1905).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fujita, H.; Li, T.R.; Yao, Y.Y. Advances in three-way decisions and granular computing. *Knowl.-Based Syst.* **2016**, *91*, 1–3. [CrossRef]
2. Yao, Y.Y. *Integrative Levels of Granularity, Human-Centric Information Processing through Granular Modelling*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 31–47.
3. Fujita, H.; Gaeta, A.; Loia, V.; Orciuoli, F. Resilience analysis of critical infrastructures: A cognitive approach based on granular computing. *IEEE Trans. Cybern.* **2018**, *49*, 1835–1848. [CrossRef] [PubMed]
4. Pedrycz, W. *Granular Computing Analysis and Design of Intelligent Systems*; CRC Press: Boca Raton, FL, USA, 2013.
5. Xu, R.; Wunsch, D.C. Survey of clustering algorithms. *IEEE Trans. Neural Netw. Learn. Syst.* **2005**, *16*, 645–678. [CrossRef] [PubMed]
6. Yang, X.B.; Qi, Y.S.; Song, X.N.; Yang, J.Y. Test cost sensitive multigranulation rough set: Model and minimal cost selection. *Inf. Sci.* **2013**, *250*, 184–199. [CrossRef]
7. Xu, W.H.; Guo, Y.T. Generalized multigranulation double-quantitative decision-theoretic rough set. *Knowl.-Based Syst.* **2016**, *105*, 190–205. [CrossRef]
8. Li, W.T.; Xu, W.H.; Zhang, X.Y.; Zhang, J. Updating approximations with dynamic objects based on local multigranulation rough sets in ordered information systems. *Artif. Intell. Rev.* **2022**, *55* , 1821–1855. [CrossRef]
9. Xu, W.H.; Yu, J.H. A novel approach to information fusion in multi-source datasets: A granular computing viewpoint. *Inf. Sci.* **2017**, *378*, 410–423. [CrossRef]
10. Chen, X.W.; Xu, W.H. Double-quantitative multigranulation rough fuzzy set based on logical operations in multi-source decision systems. *Int. J. Mach. Learn. Cybern.* **2022**, *13*, 1021–1048. [CrossRef]

11. Xu, W.H.; Yuan, K.H.; Li, W.T. Dynamic updating approximations of local generalized multigranulation neighborhood rough set. *Appl. Intell.* **2022**, *52*, 9148–9173. [CrossRef]
12. Yang, X.B.; Yao, Y.Y. Ensemble selector for attribute reduction. *Appl. Soft Comput.* **2018**, *70*, 1–11. [CrossRef]
13. Jiang, Z.H.; Yang, X.B.; Yu, H.L.; Liu, D.; Wang, P.X.; Qian, Y.H. Accelerator for multi-granularity attribute reduction. *Knowl.-Based Syst.* **2019**, *177*, 145–158. [CrossRef]
14. Li, J.Z.; Yang, X.B.; Song, X.N.; Li, J.H.; Wang, P.X.; Yu, D.-J. Neighborhood attribute reduction: A multi-criterion approach. *Int. J. Mach. Learn. Cybern.* **2019**, *10*, 731–742. [CrossRef]
15. Liu, K.Y.; Yang, X.B.; Yu, H.L.; Fujita, H.; Chen, X.J.; Liu, D. Supervised information granulation strategy for attribute reduction. *Int. J. Mach. Learn. Cybern.* **2020**, *11*, 2149–2163. [CrossRef]
16. Xu, S.P.; Yang, X.B.; Yu, H.L.; Yu, D.-J.; Yang, J.; Tsang, E.C.C. Multi-label learning with label-specific feature reduction. *Knowl.-Based Syst.* **2016**, *104*, 52–61. [CrossRef]
17. Liu, K.Y.; Yang, X.B.; Fujita, H. An efficient selector for multi-granularity attribute reduction. *Inf. Sci.* **2019**, *505*, 457–472. [CrossRef]
18. Liu, K.Y.; Yang, X.B.; Yu, H.L; Mi, J.S.; Wang, P.X.; Chen, X.J. Rough set based semi-supervised feature selection via ensemble selector. *Knowl.-Based Syst.* **2020**, *165*, 282–296. [CrossRef]
19. MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*; University of California Press: Berkeley, CA, USA, 1967; pp. 281–297.
20. Zhang, G.; Zhang, C.C.; Zhang, H.Y. Improved k-means algorithm based on density canopy. *Knowl.-Based Syst.* **2018**, *145*, 289–297. [CrossRef]
21. Wang, P.X.; Shi, H.; Yang, X.B.; Mi, J.S. Three-way k-means: Integrating k-means and three-way decision. *Int. J. Mach. Learn. Cybern.* **2019**, *10*, 2767–2777. [CrossRef]
22. Yao, Y.Y. Three-way decisions with probabilistic rough sets. *Inf. Sci.* **2010**, *180*, 341–353. [CrossRef]
23. Yao, Y.Y. The superiority of three-way decisions in probabilistic rough set models. *Inf. Sci.* **2011**, *181*, 1080–1096. [CrossRef]
24. Colorni, A.; Dorigo, M.; Maniezzo, V. Distributed 0ptimization by ant colonies. In Proceedings of the European Conference on Artificial Life, Paris, France, 11–13 December 1991.
25. Yao, Y.Y. Three-way decision and granular computing. *Int. J. Approx. Reason.* **2018**, *103*, 107–123. [CrossRef]
26. Luo, J.F.; Hu, M.J.; Qin, K.Y. Three-way decision with incomplete information based on similarity and satisfiability. *Int. J. Approx. Reason.* **2020**, *120*, 151–183. [CrossRef]
27. Xu, J.F.; Zhang, Y.J.; Miao, D.Q. Three-way confusion matrix for classification: A measure driven view. *Inf. Sci.* **2020**, *507*, 772–794. [CrossRef]
28. Qi, J.J.; Qian, T.; Wei, L. The connections between three-way and classical concept lattices. *Knowl.-Based Syst.* **2016**, *91*, 143–151. [CrossRef]
29. Li, J.H.; Huang, C.C.; Qi, J.J.; Qian, Y.H.; Liu, W.Q. Three-way cognitive concept learning via multi-granularity. *Inf. Sci.* **2017**, *378*, 244–263. [CrossRef]
30. Yuan, K.H.; Xu, W.H.; Li, W.T.; Ding, W.P. An incremental learning mechanism for object classification based on progressive fuzzy three-way concept. *Inf. Sci.* **2022**, *584*, 127–147. [CrossRef]
31. Li, X.N. Three-way fuzzy matroids and granular computing. *Int. J. Approx. Reason.* **2019**, *114*, 44–50. [CrossRef]
32. Fujita, H.; Gaeta, A.; Loia, V.; Orciuoli, F. Improving awareness in early stages of security analysis: A zone partition method based on GrC. *Appl. Intell.* **2019**, *49*, 1063–1077. [CrossRef]
33. Yang, X.; Li, T.R.; Liu, D.; Fujita, H. A temporal-spatial composite sequential approach of three-way granular computing. *Inf. Sci.* **2019**, *486*, 171–189. [CrossRef]
34. Hu, C.X.; Zhang, L. Incremental updating probabilistic neighborhood three-way regions with time-evolving attributes. *Int. J. Approx. Reason.* **2020**, *120*, 1–23. [CrossRef]
35. Yu, H. A framework of three-way cluster analysis. In Proceeding of the International Joint Conference on Rough Sets, Olsztyn, Poland, 3–7 July 2017; Volume 425, pp. 300–312.
36. Shah, A.; Azam, N.; Ali, B.; Khan, M.T.; Yao, J.T. A three-way clustering approach for novelty detection. *Inf. Sci.* **2021**, *569*, 650–668. [CrossRef]
37. Wang, P.X.; Yao, Y.Y. CE3: A three-way clustering method based on mathematical morphology. *Knowl.-Based Syst.* **2018**, *155*, 54–65. [CrossRef]
38. Jia, X.Y.; Rao, Y.; Li, W.W.; Yang, S.C.; Yu, H. An automatic three-way clustering method based on sample similarity. *Int. J. Mach. Learn. Cybern.* **2021**, *12*, 1545–1556. [CrossRef]
39. Fan, J.C.; Wang, P. X.; Jiang, C.M.; Yang, X. B.; Song, J.J. Ensemble learning using three-way density-sensitive spectral clustering. *Int. J. Approx. Reason.* **2022**, *149*, 70–84. [CrossRef]
40. Shah, A.; Azam, N.; Alanazi, E.; Yao, J.T. Image blurring and sharpening inspired three-way clustering approach. *Appl. Intell.* **2022**. [CrossRef]
41. Afridi, M.K.; Azam, N.; Yao, J.T. A three-way clustering approach for handling missing data using gtrs. *Int. J. Approx. Reason.* **2018**, *98*, 11–24. [CrossRef]
42. Wang, P.X.; Chen, X.J. Three-way ensemble clustering forincomplete data. *IEEE Access* **2020**, *8*, 91855–91864. [CrossRef]

43. Wang, P.X.; Yang, X.B. Three-way clustering method based on stability theory. *IEEE Access* **2021**, *9*, 33944–33953. [CrossRef]
44. Yu, H.; Chen, L.Y.; Yao, J.T. A three-way density peak clustering method based on evidence theory. *Knowl.-Based Syst.* **2021**, *211*, 106532. [CrossRef]
45. Fan, J.C.; Wang, X.X.; Wu, T.F.; Zhu, J.; Wang, P.X. Three-way ensemble clustering based on sample's perturbation theory. *Mathematics* **2022**, *10*, 2598. [CrossRef]
46. Wu, T.F.; Fan, J.C.; Wang, P.X. An improved three-way clustering based on ensemble strategy. *Mathematics* **2022**, *10*, 1457. [CrossRef]
47. Vinh, N.X.; Epps, J.; Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.* **2010**, *11*, 2837–2854.
48. Bezdek, J.C. Pattern recognition with fuzzy objective function algorithms. *Adv. Appl. Pattern Recognit.* **1981**, *22*, 203–239.