

Article

New Interval Improved Fuzzy Partitions Fuzzy C-Means Clustering Algorithms under Different Distance Measures for Symbolic Interval Data Analysis

Sheng-Chieh Chang ¹, Wei-Ching Chuang ¹ and Jin-Tsong Jeng ^{2,*}

¹ Department of Electro-Optical Engineering, National Formosa University, Yunlin 632, Taiwan; qooqoo331@hotmail.com (S.-C.C.); eocwc@nfu.edu.tw (W.-C.C.)

² Department of Computer Science and Information Engineering, National Formosa University, Yunlin 632, Taiwan

* Correspondence: tsong@nfu.edu.tw

Abstract: Symbolic interval data analysis (SIDA) has been successfully applied in a wide range of fields, including finance, engineering, and environmental science, making it a valuable tool for many researchers for the incorporation of uncertainty and imprecision in data, which are often present in real-world scenarios. This paper proposed the interval improved fuzzy partitions fuzzy C-means (IIFPFCM) clustering algorithm from the viewpoint of fast convergence that independently combined with Euclidean distance and city block distance. The two proposed methods both had a faster convergence speed than the traditional interval fuzzy c-means (IFCM) clustering method in SIDA. Moreover, there was a problem regarding large and small group division for symbolic interval data. The proposed methods also had better performance results than the traditional interval fuzzy c-means clustering method in this problem. In addition, the traditional IFCM clustering method will be affected by outliers. This paper also proposed the IIFPFCM algorithm to deal with outliers from the perspective of interval distance measurement. From experimental comparative analysis, the proposed IIFPFCM clustering algorithm with the city block distance measure was found to be suitable for dealing with SIDA with outliers. Finally, nine symbolic interval datasets were assessed in the experimental results. The statistical results of convergence and efficiency on performance revealed that the proposed algorithm has better results.

Keywords: symbolic interval data analysis; interval improved fuzzy partitions fuzzy C-means clustering; city block distance measure; outlier



Citation: Chang, S.-C.; Chuang, W.-C.; Jeng, J.-T. New Interval Improved Fuzzy Partitions Fuzzy C-Means Clustering Algorithms under Different Distance Measures for Symbolic Interval Data Analysis. *Appl. Sci.* **2023**, *13*, 12531. <https://doi.org/10.3390/app132212531>

Academic Editor: Alessandra Biancolillo

Received: 24 October 2023

Revised: 14 November 2023

Accepted: 14 November 2023

Published: 20 November 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Symbolic interval data analysis (SIDA) is a statistical method that allows for more efficient and intuitive analysis. In general, SIDA allows for the precise analysis of interval data by converting the data into symbolic representations. This makes computations easier and more accurate. Regarding the advantages and characteristics of SIDA, (1) SIDA can incorporate uncertainty and imprecision in data, which are often present in real-world scenarios. This makes it a useful tool for clustering in uncertain situations. (2) The symbolic representation of interval data can significantly reduce the size of the dataset, making it easier and faster to analyze. (3) SIDA can be applied to a wide range of data types and can be used for a variety of analyses, including clustering, classification, and regression. (4) SIDA is computationally efficient, making it a valuable tool for analyzing large datasets. Hence, SIDA is a useful method for analyzing interval data in a precise, intuitive, and efficient manner. Its flexibility and ability to incorporate uncertainty make it a valuable tool for fuzzy clustering in a wide range of fields [1,2]. He et al. [3] introduced an interval optimization method for applying SIDA in finance. Zhou et al. [4] presented a trajectory-recovery algorithm based on interval analysis, applied to SIDA in engineering. Yamaka et al. [5]

suggested a convex combination approach for artificial neural networks with interval data to forecast fluctuations or uncertainty in economic, financial, and environmental data. Fordellone et al. [6] proposed a maximum entropy fuzzy clustering approach for interval-valued data in cancer detection within the realm of SIDA in medicine. Freitas et al. [7] introduced a novel autocorrelation index with applications in COVID-19 and rent prices, using interval-valued data within the framework of SIDA in medicine. These studies explored various applications of SIDA across diverse fields, such as finance, engineering, environmental science, and medicine. The publication of these works underscores the significance of SIDA in addressing challenges under different domains.

As information technology rapidly advances, collecting extensive datasets has become more accessible. When dealing with substantial data, the goal is to uncover the latent information within it. Clustering analysis is a valuable tool for extracting meaningful insights from data [8,9]. In general, clustering analysis can be categorized into two types based on their modeling assumptions: probabilistic models and non-probabilistic models [10]. Methods relying on probabilistic models typically assume that the data follows a specific probability distribution and are often solved using the maximum likelihood method, with the expectation–maximization (EM) algorithm being one of the most prevalent techniques [11]. Non-probabilistic methods, on the other hand, are grounded in objective functions that aim to either maximize similarity or minimize dissimilarity to achieve optimization. These methods include hierarchical clustering [12–14] and partitioning clustering [15–17]. An important development was Bezdek’s introduction of an improved K-means clustering algorithm, known as the fuzzy C-means (FCM) clustering algorithm [18]. FCM incorporates fuzzy theory concepts, allowing data points to have membership in multiple groups rather than belonging to a single, specific cluster. Papers [19,20] proposed a modified fuzzy C-means clustering algorithm under single-valued data on image process. Papers [21–23] proposed a modified fuzzy C-means clustering algorithm under single-valued data on data analysis. Wang et al. [24] proposed guided filter-based fuzzy clustering under single-valued data for general data analysis. Papers [25–28] proposed hybrid algorithms under the fuzzy C-means algorithm with single-valued data on different applications. These studies collectively concentrate on the analysis and application of single-valued data under hybrid or modified fuzzy C-means clustering algorithms; that is, these papers all focused on single-valued data analysis and application. Frank Höppner and Frank Klawonn introduced an enhanced FCM clustering algorithm [29] known as the improved fuzzy partitions fuzzy C-means clustering (IFPFCM) algorithm. The IFPFCM algorithm encourages the assignment of crisp membership degrees by incorporating a ‘reward’ term for membership degrees close to zero or one. This modification of the objective function used in fuzzy clustering results in distinct membership functions, facilitating rapid convergence and improved performance.

A large amount of data can be expressed in intervals to simplify the amount of data and contain hidden data. Interval data represent a range of values between two bounds denoted as $[a_l, b_r]$. The a_l and b_r are called the starting point and ending point of an interval, respectively. In general, interval data can represent the analysis of data in a specific area, i.e., user data mining [30]. Interval data definition and distance measures were quickly discussed, as were their applications in clustering [31,32]. Peng W et al. [33] (2006) proposed interval data clustering with applications. In 2007, De Carvalho [34] introduced a fuzzy C-means clustering method tailored for symbolic interval data (SID). This marked a traditional application of the fuzzy C-means algorithm to address SID. Subsequently, Jeng et al. [35] proposed the interval possibilistic FCM (IPFCM) clustering method for SID, while Jeng et al. [36] developed a rough IPFCM clustering algorithm. The latter utilized derived formulas for fuzzy membership degrees (FMDs) and possibilistic membership degrees (PMDs) to improve clustering in SID, especially when dealing with noisy data and overlapping clusters. However, these methods involved complex mathematics and led to lengthy computations. The motivation behind this paper was to extend the IPFCM approach with IIFPFCM, emphasizing clear membership, which results in rapid convergence but may be

susceptible to outliers. To address this limitation, we also introduced modifications in the city block distance calculation to mitigate these challenges.

Regarding the advantages of the proposed IIFPCM, we used Figures 1 and 2 with IFCM with Euclidean distance (IFCME) and IIFPCM with Euclidean distance (IIFPCME) algorithms on the data involving five groups to show that the division of membership in IIFPCME is more crisply. The use of this concept sped up the convergence of IIFPCME in this paper.

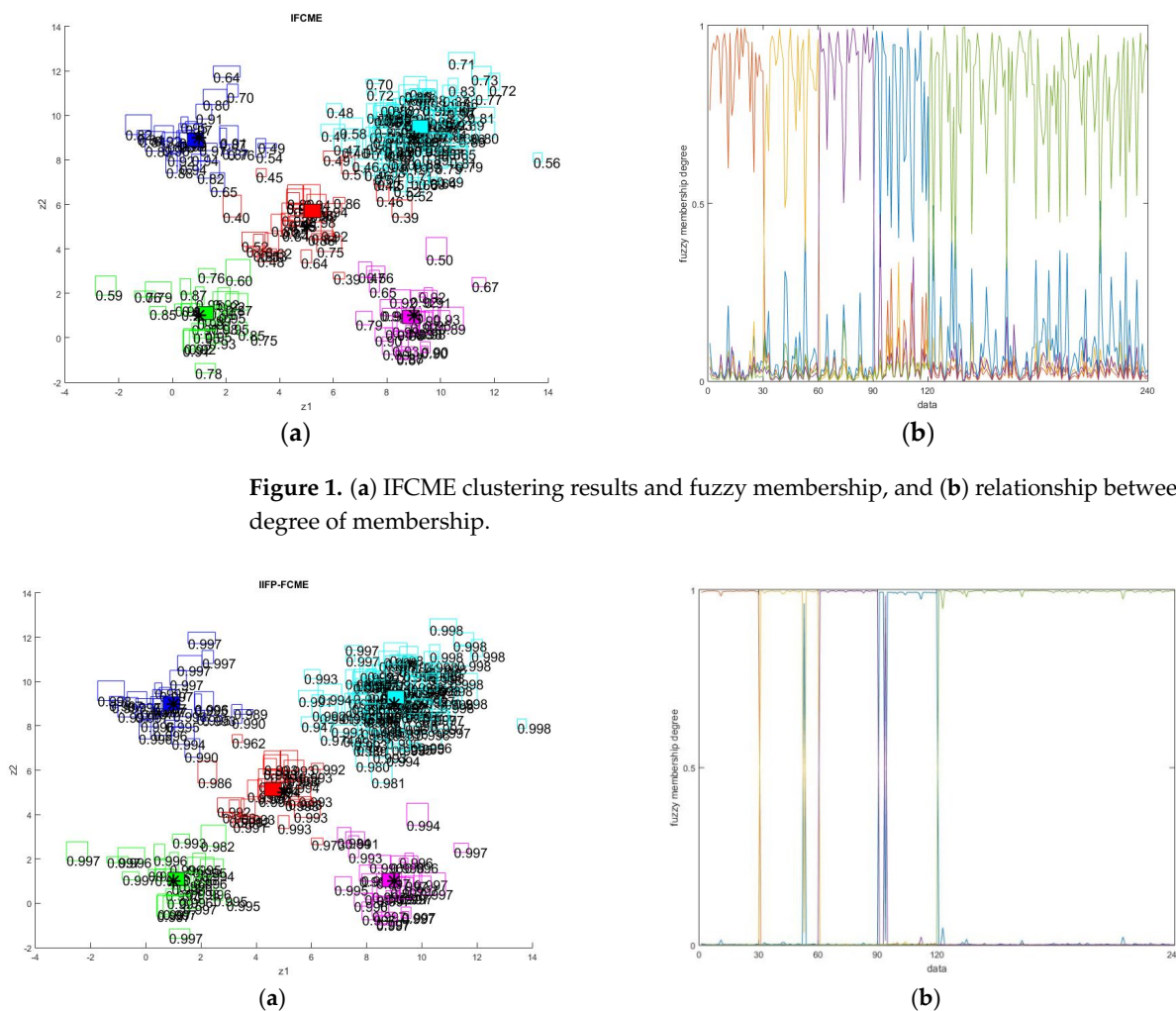


Figure 1. (a) IFCME clustering results and fuzzy membership, and (b) relationship between data and degree of membership.

Figure 2. (a) IIFPCME clustering results and fuzzy membership, and (b) relationship between data and degree of membership.

Hence, the contributions in this paper are two fuzzy clustering approaches called the new interval fuzzy partitions fuzzy C-means clustering algorithms that were proposed. The proposed algorithms independently combined Euclidean distance and city block distance to achieve fast convergence. The proposed methods showed a faster convergence speed and better performance results compared to the traditional interval fuzzy c-means clustering method in SIDA. Additionally, symbolic interval data often present a challenge of large and small group division, which was also addressed using the proposed methods. Additionally, the IIFPCM algorithm was introduced to handle outliers through interval distance measurement. The experimental comparative analysis revealed that the IIFPCM clustering algorithm with the city block distance measure was suitable for dealing with SIDA with outliers. Finally, this study tested nine groups of data, and the statistical results demonstrated that the proposed algorithm outperformed the traditional method in terms of convergence and efficiency.

The organization of the rest of this paper is as follows. In Section 2, backgrounds on definition and distance measures for interval data are detailed. In Section 3, an improved fuzzy partitions fuzzy C-means clustering algorithm is briefly introduced. In Section 4, two interval improved fuzzy partitions fuzzy C-means clustering algorithms are proposed and discussed. The simulation results are shown in Section 5. Finally, the conclusions are summarized in Section 6.

2. Definition and Distance Measures for Symbolic Interval Data

In this paper, the definition of an interval was outlined as the lower and upper bounds. If it was on the line, it was expressed as $[a, b]$, where $b > a$. When we represent data in multiple intervals, it can be expressed as the following: let $A = (A^1, \dots, A^p)$ and $B = (B^1, \dots, B^p)$ be two interval datasets, where $A^j = [a^j, b^j]$ and $B^j = [\alpha^j, \beta^j]$ indicate the values of the interval for the $j \in 1, \dots, p$ [30].

The first distance measure is termed the L1 norm, or city block distance [31]. The representation in the interval is the sum of the absolute values of the differences in all intervals:

$$D_{L1}(A, B) = \sum_{j=1}^p D_{L1}(A^j, B^j) = \sum_{j=1}^p (|a^j - \alpha^j| + |b^j - \beta^j|), \quad A = [a^j, b^j], \quad B = [\alpha^j, \beta^j], \quad j \in 1, \dots, p. \quad (1)$$

The second is called the L2 norm, or Euclidean distance [32]; the representation in the interval is the sum of the square roots of the differences in all intervals:

$$D_{L2}(A, B) = \sum_{j=1}^p D_{L2}(A^j, B^j) = \sum_{j=1}^p \left((a^j - \alpha^j)^2 + (b^j - \beta^j)^2 \right), \quad A = [a^j, b^j], \quad B = [\alpha^j, \beta^j], \quad j \in 1, \dots, p. \quad (2)$$

Figure 3 shows the relationship between city block distance and Euclidean distance under the center point in 2-dimensional space. The equidistant distance between the city block distance and the center point is similar to the rhombus in Figure 3 with the black line. Moreover, the Euclidean distance was equidistant from the center point according to the Pythagorean theorem, which is similar to a circle shape with a red line [37].

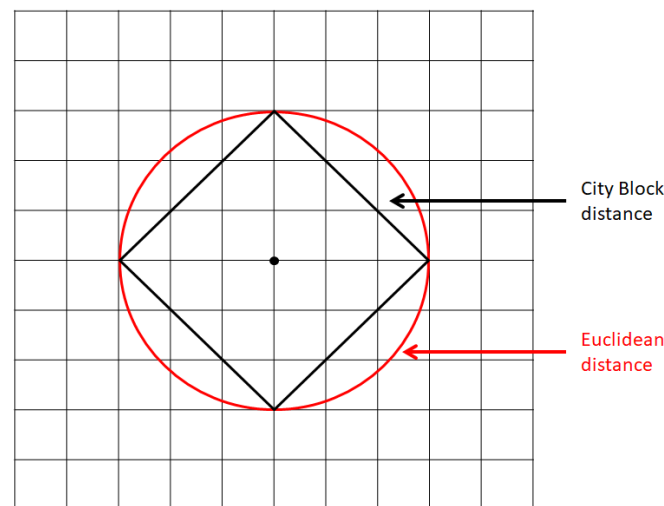


Figure 3. The relationship between city block distance and Euclidean distance and the center point in 2-dimensional space.

It can be seen from Figure 3 that different distance measures have different characteristics for different sets of data, and we discussed these differences in the experimental results section.

3. Improved Fuzzy Partitions Fuzzy C-Means Clustering Algorithm

In this section, the IFPFCM clustering algorithm has been briefly reviewed. It considers rewarding more crisp membership degrees. By adding the second term, when the degree of membership becomes fuzzier, the increase in the second term can minimize Equation (3) and use this to reward the degree of membership, making it more clearly divided. From Equation (3), a second term was evaluated as $-e_k/4$; if the membership degrees become fuzzier, the second term increases the “reward” term for membership degrees near 0 and 1 in order to force a more crisply assignment. Since we sought to minimize Equation (4), this modification rewards crisp membership degrees. The IFPFCM clustering algorithm is used in Lagrangian theory to minimize the objective function J_{ifpfc} , as shown in Equation (3), and the constraints are in Equation (4). The IFPFCM clustering algorithm has two updated equations for u_{ik} and v_i , which are outlined in Equations (5) and (6), respectively [29]:

$$J_{ifpfc}(V, U, X) = \sum_{i=1}^c \sum_{k=1}^n u_{ik}^2 d_{ik}^2 - \sum_{k=1}^n e_k \sum_{i=1}^c \left(u_{ik} - \frac{1}{2}\right)^2. \quad (3)$$

$$\text{subject to } \sum_{i=1}^c u_{ik} = 1. \quad (4)$$

$$u_{ik} = \frac{1}{\sum_{s=1}^c \left(\frac{d_{ik}^2 - e_k}{d_{sk}^2 - e_k}\right)}, \quad (5)$$

$$v_i = \frac{\sum_{k=1}^n u_{ik}^2 x_k}{\sum_{k=1}^n u_{ik}^2}. \quad (6)$$

where u_{ik} denotes the fuzzy membership degree, $U = [u_{ik}]$ is a $c \times n$ matrix, x_k is k th data of the X dataset, v_i is the i th prototype of the V prototype matrix, and d_{ik} is the distance measure between x_k and v_i . As $d_{ik}^2 - e_k$ always remains positive, the maximal reward following is denoted in the following equation:

$$e_k = \min \left\{ d_{ik}^2 - \eta \mid i \in \{1, 2, \dots, c\} \right\}, \eta \in (0, \infty). \quad (7)$$

4. Interval Improved Fuzzy Partitions Fuzzy C-Means Clustering Algorithm

The new objective function was proposed to extend the IFPFCM clustering algorithm so that interval data can be processed in this paper. The proposed method combines the advantages of the IFPFCM clustering algorithm, rewarding crisp membership degrees, and the characteristics of fast convergence. However, the proposed method under Euclidean distance will still be affected by outliers. In order to overcome this shortcoming, adding interval city block distance to the proposed method overcame this problem. The derivation of the IIFPFCM clustering algorithm under Euclidean distance has been proposed in Section 4.1. The derivation of the IIFPFCM clustering algorithm under city block distance has been proposed in Section 4.2.

4.1. Interval Improved Fuzzy Partitions Fuzzy C-Means under Euclidean Distance

The new objective function, $J_{iifpfcme}$, for the IIFPFCM clustering algorithm is based on a squared Euclidean distance between the vectors of intervals given in Equation (8) and of which the constraints are given in Equation (9):

$$\begin{aligned} J_{iifpfcme}(\bar{G}, \bar{U}, X) &= \sum_{i=1}^c \sum_{k=1}^n \bar{u}_{ik}^2 D_{L2}(\vec{x}_k, \vec{g}_i) - \sum_{k=1}^n \bar{e}_k \sum_{i=1}^c \left(\bar{u}_{ik} - \frac{1}{2}\right)^2, \\ &= \sum_{i=1}^c \sum_{k=1}^n \bar{u}_{ik}^2 \sum_{j=1}^p \left[(a_k^j - \bar{\alpha}_i^j)^2 + (b_k^j - \bar{\beta}_i^j)^2 \right] - \sum_{k=1}^n \bar{e}_k \sum_{i=1}^c \left(\bar{u}_{ik} - \frac{1}{2}\right)^2, \end{aligned} \quad (8)$$

$$\text{subject to } \sum_{i=1}^c \bar{u}_{ik} = 1. \quad (9)$$

In Equation (8), $\bar{u}_{ik}$ is the fuzzy membership degree of pattern k in the i th cluster, and $\bar{U} = [\bar{u}_{ik}]$ is a $c \times n$ matrix; the $\bar{e}_k$ is similar to the IFPFCM defined in Equation (10):

$$\bar{e}_k = \min \left\{ \sum_{j=1}^p \left[(a_k^j - \bar{\alpha}_i^j)^2 + (b_k^j - \bar{\beta}_i^j)^2 \right] - \eta \mid i \in \{1, 2, \dots, c\} \right\}, \eta \in (0, \infty). \quad (10)$$

Using the Lagrange multipliers method for the objective function $J_{iifpfcme}$, the nonlinear unconstraint optimization can be obtained as:

$$L_{iifpfcme} = \sum_{i=1}^c \sum_{k=1}^n \bar{u}_{ik}^2 \sum_{j=1}^p \left[(a_k^j - \bar{\alpha}_i^j)^2 + (b_k^j - \bar{\beta}_i^j)^2 \right] - \sum_{k=1}^n \bar{e}_k \sum_{i=1}^c \left(\bar{u}_{ik} - \frac{1}{2} \right)^2 - \sum_{k=1}^n \lambda_k \left(\sum_{i=1}^c \bar{u}_{ik} - 1 \right). \quad (11)$$

The $\bar{u}_{ik}$ was considered for the minimum problem; to derive an updated equation for the fuzzy membership degree, $\bar{u}_{ik}$ can be solved as follows:

$$\frac{\partial L_{iifpfcme}}{\partial \bar{u}_{ik}} = 0 \Leftrightarrow 2\bar{u}_{ik} \sum_{j=1}^p \left[(a_k^j - \bar{\alpha}_i^j)^2 + (b_k^j - \bar{\beta}_i^j)^2 \right] - 2\bar{e}_k \left(\bar{u}_{ik} - \frac{1}{2} \right) - \lambda_k = 0, \quad (12)$$

for $i = 1, \dots, c$ and $k = 1, \dots, n$. Thus, Equation (12) can be rewritten as:

$$\bar{u}_{ik} = \frac{\lambda_k - \bar{e}_k}{2 \left(\sum_{j=1}^p \left[(a_k^j - \bar{\alpha}_i^j)^2 + (b_k^j - \bar{\beta}_i^j)^2 \right] - \bar{e}_k \right)}, \quad (13)$$

for $i = 1, \dots, c$ and $k = 1, \dots, n$. Substituting Equation (13) into Equation (9) is obtained as:

$$\sum_{s=1}^c \frac{\lambda_k - \bar{e}_k}{2 \left(\sum_{j=1}^p \left[(a_k^j - \bar{\alpha}_s^j)^2 + (b_k^j - \bar{\beta}_s^j)^2 \right] - \bar{e}_k \right)} = 1. \quad (14)$$

According to Equations (13) and (14), the updated equation for the fuzzy membership degree $\bar{u}_{ik}$ is:

$$\bar{u}_{ik} = \frac{1}{\sum_{s=1}^c \left(\frac{\sum_{j=1}^p \left[(a_k^j - \bar{\alpha}_i^j)^2 + (b_k^j - \bar{\beta}_i^j)^2 \right] - \bar{e}_k}{\sum_{j=1}^p \left[(a_k^j - \bar{\alpha}_s^j)^2 + (b_k^j - \bar{\beta}_s^j)^2 \right] - \bar{e}_k} \right)}, \quad (15)$$

for $i = 1, \dots, c$ and $k = 1, \dots, n$.

The prototype $\vec{g}_i = (\bar{g}_i^1, \dots, \bar{g}_i^p)$ of the i th cluster, which minimizes the objective function $J_{iifpfcme}$, has the bounds of the interval $\bar{g}_i^j = [\bar{\alpha}_i^j, \bar{\beta}_i^j]$ ($j = 1, \dots, p$). Hence, the updated formula for the prototype $\bar{g}_i^j$ accords to the following:

$$\frac{\partial L_{iifpfcme}}{\partial \bar{\alpha}_i^j} = 0 \Leftrightarrow 2 \sum_{k=1}^n \bar{u}_{ik}^2 (a_k^j - \bar{\alpha}_i^j) = 0 \text{ and } \frac{\partial L_{iifpfcme}}{\partial \bar{\beta}_i^j} = 0 \rightarrow 2 \sum_{k=1}^n \bar{u}_{ik}^2 (b_k^j - \bar{\beta}_i^j) = 0, \quad (16)$$

for $i = 1, \dots, c$ and $j = 1, \dots, p$. Thus, Equation (16) can be rewritten as:

$$\bar{\alpha}_i^j = \frac{\sum_{k=1}^n \bar{u}_{ik}^2 a_k^j}{\sum_{k=1}^n \bar{u}_{ik}^2} \text{ and } \bar{\beta}_i^j = \frac{\sum_{k=1}^n \bar{u}_{ik}^2 b_k^j}{\sum_{k=1}^n \bar{u}_{ik}^2}. \quad (17)$$

In Equation (8), adding a second term evaluates to $-e_k/4$. When membership degrees become fuzzier, the second term increases as a ‘reward’ for membership degrees near 0 and 1, aiming to enforce a crisper assignment. The adoption of this concept accelerates the convergence of IIFPFCME in this study. By introducing the second term, as membership degrees become fuzzier, its increase can minimize Equation (8). This effectively rewards the degree of membership, resulting in clearer divisions.

4.2. Interval Improved Fuzzy Partitions Fuzzy C-Means under City Block Distance

The new objective function, $J_{iifpfcme}$, for the IIFPFCM clustering algorithm is based on a city block distance between the vectors of intervals given in Equation (18) and of which the constraints are given in Equation (19):

$$\begin{aligned} J_{iifpfcme}(\tilde{G}, \tilde{U}, X) &= \sum_{i=1}^c \sum_{k=1}^n \tilde{u}_{ik}^2 D_{L2}(\vec{x}_k, \vec{\tilde{g}}_i) - \sum_{k=1}^n e_k^c \sum_{i=1}^c \left(\tilde{u}_{ik} - \frac{1}{2}\right)^2, \\ &= \sum_{i=1}^c \sum_{k=1}^n \tilde{u}_{ik}^2 \sum_{j=1}^p \left(|a_k^j - \tilde{\alpha}_i^j| + |b_k^j - \tilde{\beta}_i^j|\right) - \sum_{k=1}^n \tilde{e}_k \sum_{i=1}^c \left(\tilde{u}_{ik} - \frac{1}{2}\right)^2, \end{aligned} \quad (18)$$

$$\text{subject to } \sum_{i=1}^c \tilde{u}_{ik} = 1. \quad (19)$$

In Equation (18), $\tilde{u}_{ik}$ is the fuzzy membership degree of pattern k in the i th cluster, and $\tilde{U} = [\tilde{u}_{ik}]$ is a $c \times n$ matrix; the $\tilde{e}_k$ is similar to the IFPFCM defined in Equation (20):

$$\tilde{e}_k = \min \left\{ \sum_{j=1}^p \left(|a_k^j - \tilde{\alpha}_i^j| + |b_k^j - \tilde{\beta}_i^j|\right) - \eta \mid i \in \{1, 2, \dots, c\} \right\}, \eta \in (0, \infty). \quad (20)$$

Using the Lagrange multipliers method for the objective function $J_{iifpfcme}$, the nonlinear unconstraint optimization can be obtained as:

$$L_{iifpfcme} = \sum_{i=1}^c \sum_{k=1}^n \tilde{u}_{ik}^2 \sum_{j=1}^p \left(|a_k^j - \tilde{\alpha}_i^j| + |b_k^j - \tilde{\beta}_i^j|\right) - \sum_{k=1}^n \tilde{e}_k \sum_{i=1}^c \left(\tilde{u}_{ik} - \frac{1}{2}\right)^2 - \sum_{k=1}^n \lambda_k \left(\sum_{i=1}^c \tilde{u}_{ik} - 1\right). \quad (21)$$

The $\tilde{u}_{ik}$ was considered for the minimum problem; to obtain an updated equation of the fuzzy membership degree, u_{ik} can be solved as followed:

$$\frac{\partial L_{iifpfcme}}{\partial \tilde{u}_{ik}} = 0 \Leftrightarrow 2\tilde{u}_{ik} \sum_{j=1}^p \left(|a_k^j - \tilde{\alpha}_i^j| + |b_k^j - \tilde{\beta}_i^j|\right) - 2\tilde{e}_k \left(\tilde{u}_{ik} - \frac{1}{2}\right) - \lambda_k = 0, \quad (22)$$

for $i = 1, \dots, c$ and $k = 1, \dots, n$. Thus, Equation (22) can be rewritten as:

$$\tilde{u}_{ik} = \frac{\lambda_k - \tilde{e}_k}{2 \left(\sum_{j=1}^p \left(|a_k^j - \tilde{\alpha}_i^j| + |b_k^j - \tilde{\beta}_i^j|\right) - \tilde{e}_k \right)}, \quad (23)$$

for $i = 1, \dots, c$ and $k = 1, \dots, n$. Substituting Equation (23) into Equation (19) is obtained as:

$$\sum_{s=1}^c \frac{\lambda_k - \tilde{e}_k}{2 \left(\sum_{j=1}^p \left(|a_k^j - \tilde{\alpha}_s^j| + |b_k^j - \tilde{\beta}_s^j|\right) - \tilde{e}_k \right)} = 1. \quad (24)$$

According to Equations (23) and (24), the updated equation for the fuzzy membership degree $\tilde{u}_{ik}$ is:

$$\tilde{u}_{ik} = \frac{1}{\sum_{s=1}^c \left(\frac{\sum_{j=1}^p (|a_k^j - \tilde{\alpha}_i^j| + |b_k^j - \tilde{\beta}_i^j|) - \tilde{e}_k}{\sum_{j=1}^p (|a_k^j - \tilde{\alpha}_s^j| + |b_k^j - \tilde{\beta}_s^j|) - \tilde{e}_k} \right)}, \quad (25)$$

for $i = 1, \dots, c$ and $k = 1, \dots, n$.

The prototype $\tilde{g}_i = (\tilde{g}_i^1, \dots, \tilde{g}_i^p)$ of the i th cluster, which minimizes the objective function $J_{iifpfcme}(\tilde{G}, \tilde{U}, X)$, has the bounds of the interval $\tilde{g}_i^j = [\tilde{\alpha}_i^j, \tilde{\beta}_i^j]$ ($j = 1, \dots, p$), which was updated according to the following problem:

$$\text{Minimize}_{\tilde{\alpha}_i^j, \tilde{\beta}_i^j} \sum_{i=1}^c \sum_{k=1}^n \tilde{u}_{ik}^2 \sum_{j=1}^p (|a_k^j - \tilde{\alpha}_i^j| + |b_k^j - \tilde{\beta}_i^j|). \quad (26)$$

This yields two well-known minimization problems in L_1 norm, such as:

$$\text{Minimize}_{\tilde{\alpha}_i^j} \sum_{i=1}^c \sum_{k=1}^n \tilde{u}_{ik}^2 \sum_{j=1}^p |a_k^j - \tilde{\alpha}_i^j|, \text{ and } \text{Minimize}_{\tilde{\beta}_i^j} \sum_{i=1}^c \sum_{k=1}^n \tilde{u}_{ik}^2 \sum_{j=1}^p |b_k^j - \tilde{\beta}_i^j|. \quad (27)$$

In De Carvalho et al., 2006 [32], the solutions of minimized $\sum_{k=1}^n |h_k^j - z_i^j|$ for z_i^j have been shown as the median of set $\{h_k^j, \text{ for } k = 1, \dots, n\}$. In De Carvalho et al., 2006 [32], there was no $\tilde{u}_{ik}^2$ in the objection function. However, in this paper, we needed to consider $\tilde{u}_{ik}^2$. Hence, we needed to consider that if $\tilde{u}_{ik} > 1/C$, then $\tilde{u}_{ik} = 1$, and if $\tilde{u}_{ik} \leq 1/C$, then $\tilde{u}_{ik} = 0$. This criteria for the fuzzy membership degree represented by it is greater than the average of all groups, so the degree $\tilde{u}_{ik}$ is strengthened, making its grouping more obvious. As a result, the updated equations of $\tilde{\alpha}_i^j$ and $\tilde{\beta}_i^j$ are obtained as:

$$\tilde{\alpha}_i^j = \tilde{a}_k^j \text{ and } \tilde{\beta}_i^j = \tilde{b}_k^j \quad (28)$$

where $\tilde{a}_k^j$ and $\tilde{b}_k^j$ denote the medians of sets $\{a_k^j, \text{ for } k = 1, \dots, n \text{ and } \tilde{u}_{ik} = 1\}$ and $\{b_k^j, \text{ for } k = 1, \dots, n \text{ and } \tilde{u}_{ik} = 1\}$, respectively.

The proposed IIFPFCM under city block distance (IIFPFCMC) clustering algorithm used the median instead of the average to update the center point of the group. Hence, the characteristic of the median is that if there are extreme values such as outliers, it can better centralize the trend, and the average will affect affected by extreme values. Hence, when the interval data have outliers, the proposed IIFPFCMC clustering algorithm can effectively avoid being affected at the center point of the update group; that is, the proposed IIFPFCMC used Equations (26)–(28) to update the center point of the cluster and used the median value instead of the mean value. Usually, the characteristic of the median is that if there are extreme values, it can better centralize the tendency. However, the average will be affected by the extreme values. Therefore, when the data have outliers, using the median calculation at the center point of the update group can effective avoid the influence of outliers. Usually, the proposed IIFPFCMC clustering algorithm directly using the median is easily affected by the center point of the initial value, resulting in the situation that the center point of the cluster appears in the same group. Therefore, we separated the initial group center points with IIFPFCME, and then used the algorithm of the median for the IIFPFCMC clustering algorithm, which makes it easier to have a good convergence effect with modified Equations (26)–(28) in this paper. Algorithm 1 presents the procedures of the proposed IIFPFCME clustering algorithm with Euclidean distance. Algorithm 2 presents the procedures of the proposed IIFPFCMC clustering algorithm with city block distance.

Algorithm 1: Procedures of the proposed IIFPFCME clustering algorithm in the following steps

Step 1: Initialization fix $c, 2 \leq c < n$; fix $\eta, 0 \leq \eta < \infty$; fix $\varepsilon > 0$;
 Set iteration counter $l = 1$ and iteration limit L ;
 Initialization $\bar{U}$.
 Step 2: Update the $\bar{\alpha}_i^l$ and $\bar{\beta}_i^l$.
 Step 3: Update the $\bar{e}_k$.
 Step 4: Update the $\bar{u}_{ik}$.
 Step 5: **if** ($|J_{iifpfcme}^{l-1} - J_{iifpfcme}^l| \leq \varepsilon$ or $l > L$) **Stop**.
Else $l = l + 1$ and go to **Step 2**.

Algorithm 2: Procedures of the proposed IIFPFCMC clustering algorithm in the following steps

Step 1: Initialization fix $c, 2 \leq c < n$; fix $\eta, 0 \leq \eta < \infty$; fix $\varepsilon > 0$;
 Set iteration counter $l = 1$ and iteration limit L ;
 Initialization $\tilde{U}$.
 Step 2: Update the $\tilde{\alpha}_i^l$ and $\tilde{\beta}_i^l$.
 Step 3: Update the $\tilde{e}_k$.
 Step 4: Update the $\tilde{u}_{ik}$.
 Step 5: **If** ($|J_{iifpfcmc}^{l-1} - J_{iifpfcmc}^l| \leq \varepsilon$ or $l > L$) **Stop**.
Else $l = l + 1$ and go to **Step 2**.

5. Simulation Results

We compared five approaches, namely IFCME, IFCMC, IPFCME, IIFPFCME, and IIFPFCMC with IDS1–2 datasets and IFCME, IFCMC, IIFPFCME, and IIFPFCMC with IDS1–6 datasets that have several datasets that include outlier interference. We also presented two real datasets concerning exchange rates for validation of the proposed methods. For a detailed description of the datasets used in this paper, please refer to [34,35]. For all datasets, we used the following computational protocols: stop requirement $\varepsilon = 0.00001$, and iteration limit $L = 300$. All experiments were run on a computer with Intel(R) Xeon(R) CPU E3-1231 v3 @ 3.40 GH, 32 GB memory, Microsoft Windows 10, and MATLAB 2018b. The performance using the root-mean-square error (RMSE) to computer the true cluster center and corresponding cluster center is defined as:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^c (g_i - \tilde{g}_i)^2}{c}} \quad (29)$$

where g_i is the true cluster center and $\tilde{g}_i$ is the corresponding cluster center.

The first dataset (DS1) consists of five clusters in 2D space which satisfy the standard normal distributions, whose centers are (1, 1), (1, 9), (9, 1), (5, 5), and (9, 9). For the first four clusters, the number of data points per cluster was 30, and their respective covariance matrices were $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. For the last cluster centered at (9, 9), there were 120 data points, and their covariance matrix was $\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$. These data are shown in Figure 4a. In order to build the interval data (symbolic) sets, this was defined as $([z_1 - \gamma_1/2, z_1 + \gamma_1/2], [z_2 - \gamma_2/2, z_2 + \gamma_2/2])$, from which γ_1 and γ_2 were drawn randomly from [0.2, 1], as shown in Figure 4b.

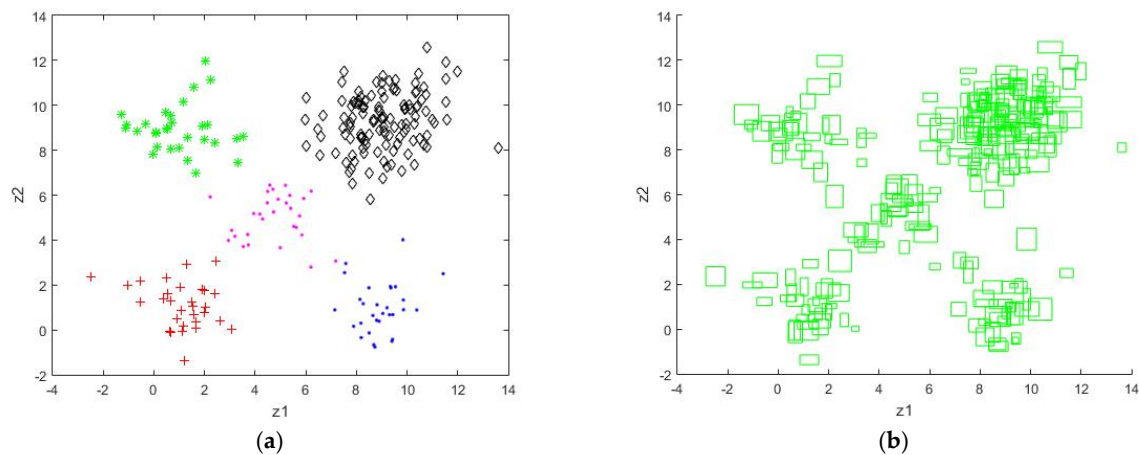


Figure 4. (a) The first dataset (DS1) and (b) interval datasets of DS1 (IDS1).

The results of IFCME and IFCMC with $m = 2$ for IDS1 are shown in Figure 5, where * is the true cluster center; the same color denotes the same clustering (group), and the symbol data with a fill-in color on each color (clustering) is the cluster center of results. The results of IIFPFCME and IIFPFCMC with parameters $\eta = 0.01$ and $\eta = 0.1$ for IDS1 are shown in Figures 6 and 7, respectively. The results of IPFCME for IDS1 are shown in Figure 8. There are three indexes: the RMSE, the average usage time of ten times, and the average of ten iterations as a comparison on Tables 1–6. The results of the comparison are shown in Table 1. From Table 1, the proposed IIFPFCME and IIFPFCMC on performance and average times all are better than IFCME, IFCMC, and IPFCME. From Table 1, the proposed IIFPFCME appears to be better than the proposed IIFPFCMC under IDS1. IDS1 is a symbolic dataset without outliers. From Figure 5, the clustering results are not good, as the center of the five clusters cannot be found correctly. Hence, we used “---” as the RMSE in Table 1. From Table 1, IPFCME takes more time, but its clustering results are not better than IIFPFCME.

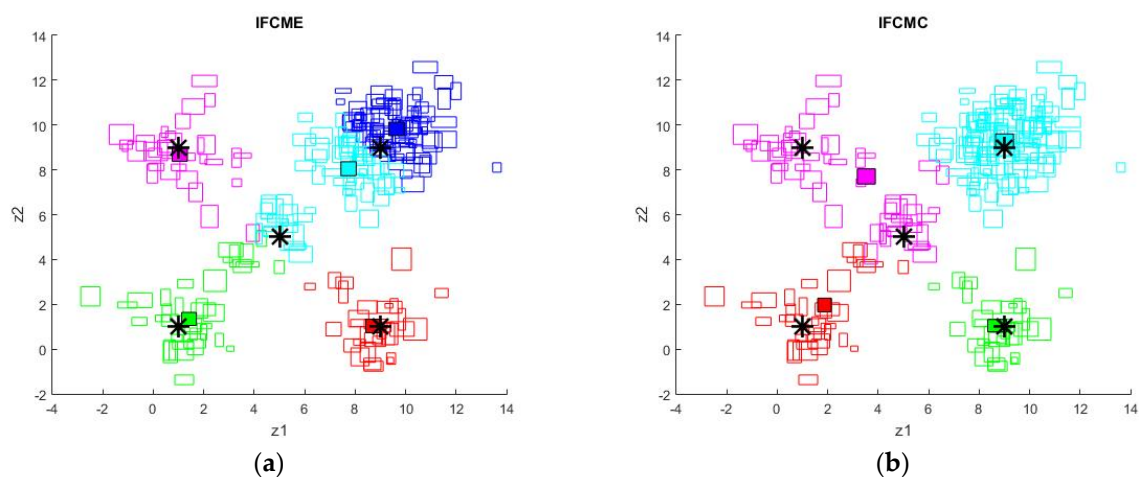


Figure 5. The results of the (a) IFCME and (b) IFCMC parameters with $m = 2$ for IDS1.

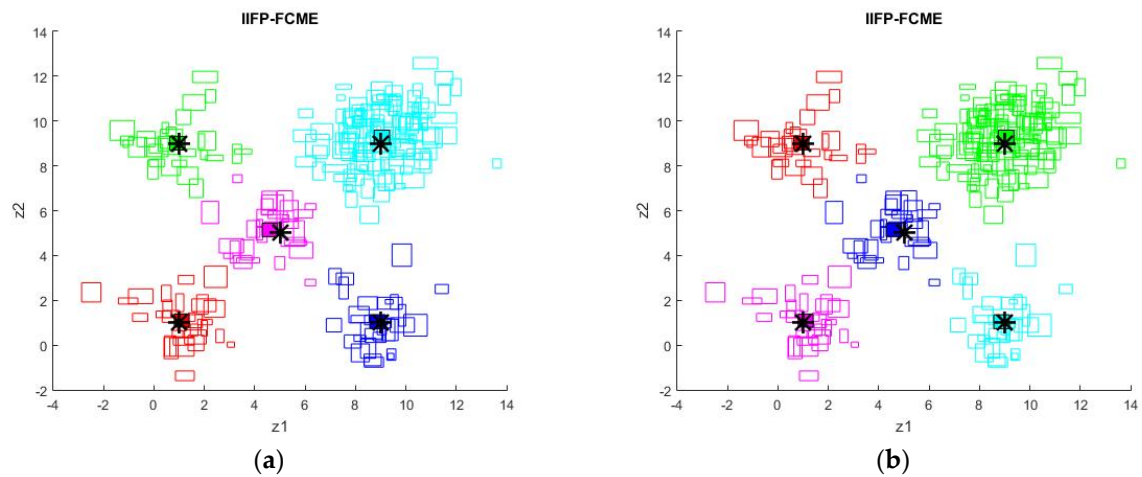


Figure 6. The results of the IIFPFCME parameter with (a) $\eta = 0.01$ and (b) $\eta = 0.1$ for IDS1.

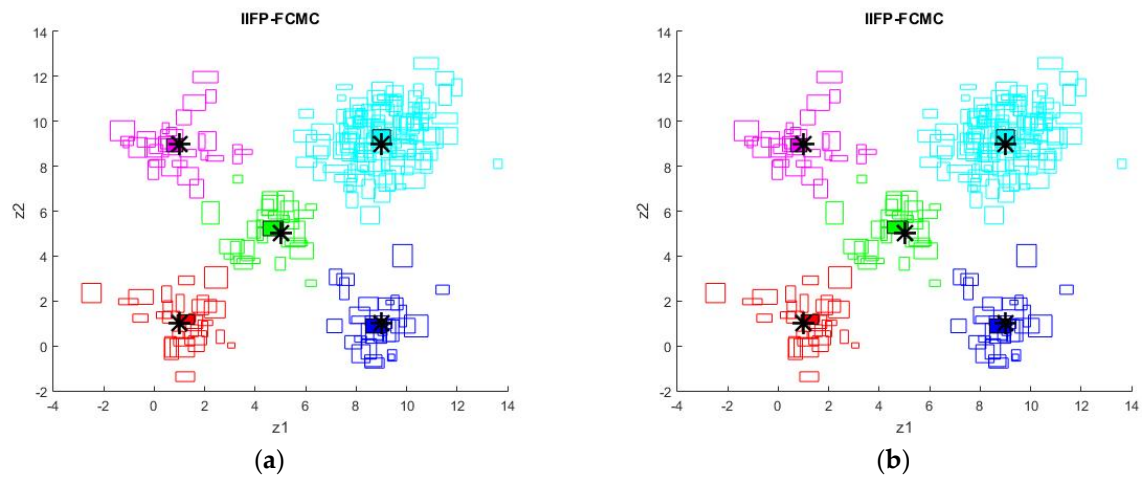


Figure 7. The results of the IIFPFCMC parameter with (a) $\eta = 0.01$ and (b) $\eta = 0.1$ for IDS1.

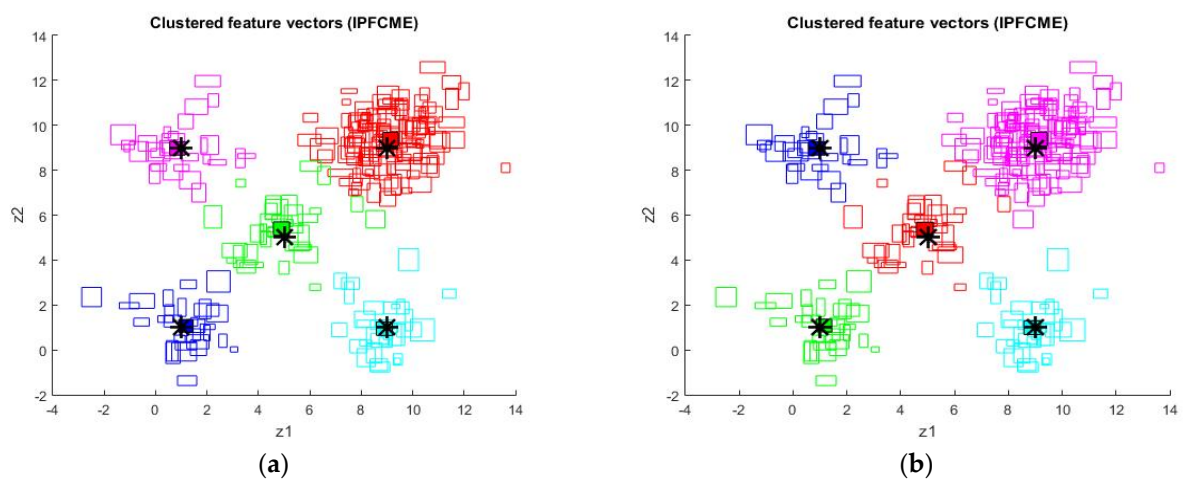


Figure 8. The results of the IPFCME parameter with (a) $m = 1.5, n = 2, a = 20, b = 50$ and (b) $m = 1.5, n = 2, a = 50, b = 50$ for IDS1.

Table 1. The results of the comparison for IDS1 with different approaches: (a) using IFCME and IIFPFCME under different η ; (b) using IFCMC and IIFPFCMC under different η ; and (c) using IPFCME.

(a)				
Compared Index/Method	IFCME ($m = 2$)	IIFPFCME ($\eta = 0.01$)	IIFPFCME ($\eta = 0.1$)	IIFPFCME ($\eta = 1$)
RMSE	---	0.2470	0.2470	0.2513
Average times on ten (sec.)	0.007468	0.004497	0.003814	0.00735
Average of ten iterations (l)	42.8	23.2	19.4	39.2
(b)				
Compared Index/Method	IFCMC ($m = 2$)	IIFPFCMC ($\eta = 0.01$)	IIFPFCMC ($\eta = 0.1$)	IIFPFCMC ($\eta = 1$)
RMSE	---	0.2924	0.2964	0.2955
Average times on ten (sec.)	0.011596	0.005645	0.006696	0.006356
Average number of ten iterations (l)	18.5	11	12	11.7
(c)				
Compared Index/Method	IPFCME ($m = 1.5, n = 2,$ $a = 20, b = 50$)		IPFCME ($m = 1.5, n = 2,$ $a = 50, b = 50$)	
RMSE	0.2885		0.2733	
Average times on ten (sec.)	0.01856		0.019446	
Average number of ten iterations (l)	35.6		37.5	

Table 2. The results of the comparison for IDS2 with different approaches: (a) using IFCME and IIFPFCME under different η ; (b) using IFCMC and IIFPFCMC under different η ; and (c) using IPFCME.

(a)				
Compared Index/Method	IFCME ($m = 2$)	IIFPFCME ($\eta = 0.01$)	IIFPFCME ($\eta = 0.01$)	IIFPFCME ($\eta = 1$)
RMSE	---	0.3124	0.3124	0.3159
Average times on ten (sec.)	0.007727	0.004025	0.005414	0.006802
Average number of ten iterations (l)	42.5	20.1	27.8	34.6
(b)				
Compared Index/Method	IFCMC ($m = 2$)	IIFPFCMC ($\eta = 0.01$)	IIFPFCMC ($\eta = 0.01$)	IIFPFCMC ($\eta = 1$)
RMSE.	---	0.273	0.2773	0.2736
Average times on ten (sec.)	0.010524	0.008881	0.007854	0.006761
Average number of ten iterations (l)	16.8	12.8	11.6	11.3
(c)				
Compared Index/Method	IPFCME ($m = 1.5, n = 2,$ $a = 20, b = 50$)		IPFCME ($m = 1.5, n = 2,$ $a = 50, b = 50$)	
RMSE	0.3545		0.3505	
Average times on ten (sec.)	0.023422		0.020567	
Average number of ten iterations (l)	44.1		38.4	

Table 3. The results of the comparison for IDS3 with different approaches: (a) using IFCME and IIFPFCME under different η ; and (b) using IFCMC and IIFPFCMC under different η .

(a)				
Compared Index/Method	IFCME ($m = 2$)	IIFPFCME ($\eta = 0.01$)	IIFPFCME ($\eta = 0.1$)	IIFPFCME ($\eta = 1$)
RMSE	---	0.4437	0.4428	0.4616
Average times on ten (sec.)	0.007077	0.003865	0.006962	0.00694
Average number of ten iterations (l)	39.6	18.9	35.2	34.7
(b)				
Compared Index/Method	IFCMC ($m = 2$)	IIFPFCMC ($\eta = 0.01$)	IIFPFCMC ($\eta = 0.1$)	IIFPFCMC ($\eta = 1$)
RMSE	---	0.3449	0.3484	0.3356
Average times on ten (sec.)	0.009187	0.007241	0.00958	0.005183
Average number of ten iterations (l)	15.1	10.5	14.3	8.9

Table 4. The results of the comparison for IDS4 with different approaches: (a) using IFCME and IIFPFCME under different η ; and (b) using IFCMC and IIFPFCMC under different η .

(a)				
Compared Index/Method	IFCME ($m = 2$)	IIFPFCME ($\eta = 0.01$)	IIFPFCME ($\eta = 0.1$)	IIFPFCME ($\eta = 1$)
RMSE	0.1668	2.7094×10^{-4}	0.0027	0.0233
Average times on ten (sec.)	0.005133	0.00203	0.001978	0.002283
Average number of ten iterations (l)	28.7	8.9	9.1	10.9
(b)				
Compared Index/Method	IFCMC ($m = 2$)	IIFPFCMC ($\eta = 0.01$)	IIFPFCMC ($\eta = 0.1$)	IIFPFCMC ($\eta = 1$)
RMSE	0.6604	0	0	0
Average times on ten (sec.)	0.008365	0.006414	0.007366	0.006735
Average number of ten iterations (l)	13.3	8.5	10	9.2

Table 5. The results of the comparison for IDS5 with different approaches: (a) using IFCME and IIFPFCME under different η ; and (b) using IFCMC and IIFPFCMC under different η .

(a)				
Compared Index/Method	IFCME ($m = 2$)	IIFPFCME ($\eta = 0.01$)	IIFPFCME ($\eta = 0.1$)	IIFPFCME ($\eta = 1$)
RMSE	0.1871	0.1778	0.1780	0.1810
Average times on ten (sec.)	0.004011	0.001845	0.002007	0.002452
Average number of ten iterations (l)	22.1	8.4	9.3	11.7
(b)				
Compared Index/Method	IFCMC ($m = 2$)	IIFPFCMC ($\eta = 0.01$)	IIFPFCMC ($\eta = 0.1$)	IIFPFCMC ($\eta = 1$)
RMSE	0.6437	0.0522	0.0522	0.0522
Average times on ten (sec.)	0.008791	0.006091	0.005886	0.006728
Average number of ten iterations (l)	12.6	8.1	7.8	9

Table 6. The results of the comparison for IDS6 with different approaches: (a) using IFCME and IIFPFCME under different η ; and (b) using IFCMC and IIFPFCMC under different η .

(a)				
Compared Index/Method	IFCME ($m = 2$)	IIFPFCME ($\eta = 0.01$)	IIFPFCME ($\eta = 0.1$)	IIFPFCME ($\eta = 1$)
RMSE	0.1801	0.2549	0.2555	0.2604
Average times on ten (sec.)	0.008623	0.003292	0.00311	0.004102
Average number of ten iterations (l)	26.7	10.3	10.8	16.5
(b)				
Compared Index/Method	IFCMC ($m = 2$)	IIFPFCMC ($\eta = 0.01$)	IIFPFCMC ($\eta = 0.1$)	IIFPFCMC ($\eta = 1$)
RMSE	0.6522	0.0626	0.0626	0.0626
Average times on ten (sec.)	0.007367	0.006087	0.007714	0.006814
Average number of ten iterations (l)	10	7.6	9.8	8.6

The second dataset (DS2) with 243 points was constructed using DS1 and another one-point (14, 4) outlier, as shown in Figure 9a. On the other hand, the interval datasets of DS2 (IDS2) are shown in Figure 9b. In this case, the symbolic data with outliers were illustrated as powerful via the proposed IIFPFCMC method.

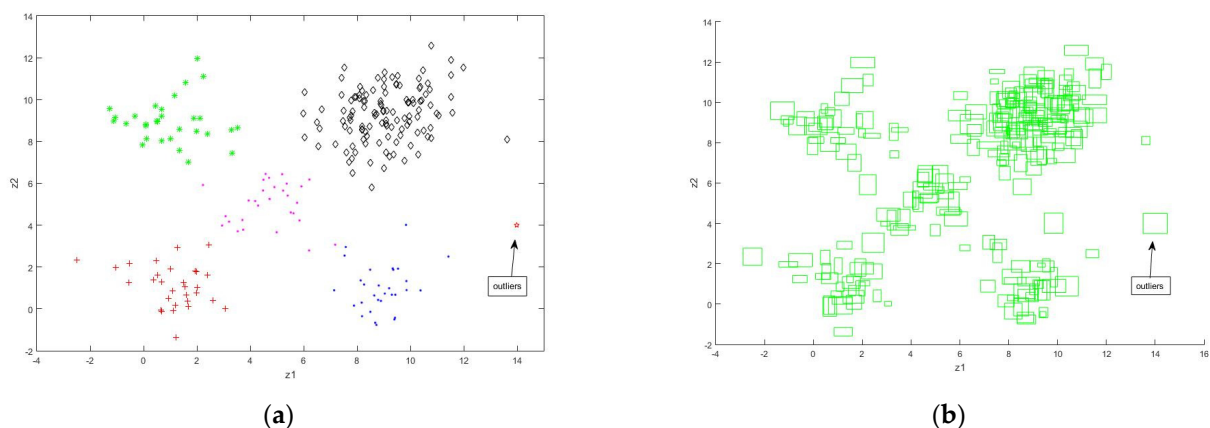


Figure 9. (a) The second dataset (DS2) and (b) interval datasets of DS2 (IDS2).

The results of IFCME and IFCMC with $m = 2$ for IDS2 are shown in Figure 10. The results of IIFPFCME and IIFPFCMC with parameters $\eta = 0.01$ and $\eta = 0.1$ for IDS2 are shown in Figures 11 and 12, respectively. The results of IPFCME for IDS2 are shown in Figure 13. The results of the comparison are shown in Table 2. From Table 2, the proposed IIFPFCME and IIFPFCMC on performance and average times are all better than IFCME, IFCMC, and IPFCME. From Table 2, the proposed IIFPFCMC is better than the proposed IIFPFCME under a symbolic set with outliers. From Figure 10, the clustering results are not good, as the center of the five clusters cannot be found correctly. Hence, we used “---” as the RMSE in Table 2. From Table 2, IPFCME takes more time, but its clustering results are not better than IIFPFCME.

The third dataset (DS3) with 252 points was constructed using DS1 and added four points, namely (14, 6), (6, 14), (6, −4), and (−4, 6), as shown in Figure 14a. On the other hand, the interval datasets of DS3 (IDS3) are shown in Figure 14b. In this case, the symbolic data with more outliers and different positions were illustrated as powerful via the proposed IIFPFCMC method. Additionally, due to IPFCME in Tables 1 and 2 on the RMSE and

average times, the results were worse than the proposed methods. Therefore, the following examples were not presented in the table.

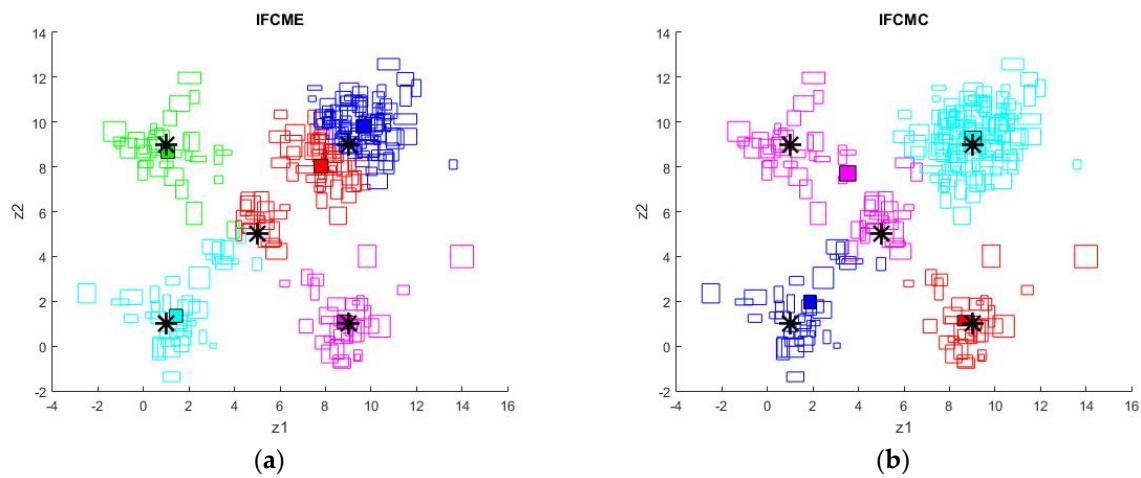


Figure 10. The results of the (a) IFCME and (b) IFCMC parameters with $m = 2$ for IDS2.

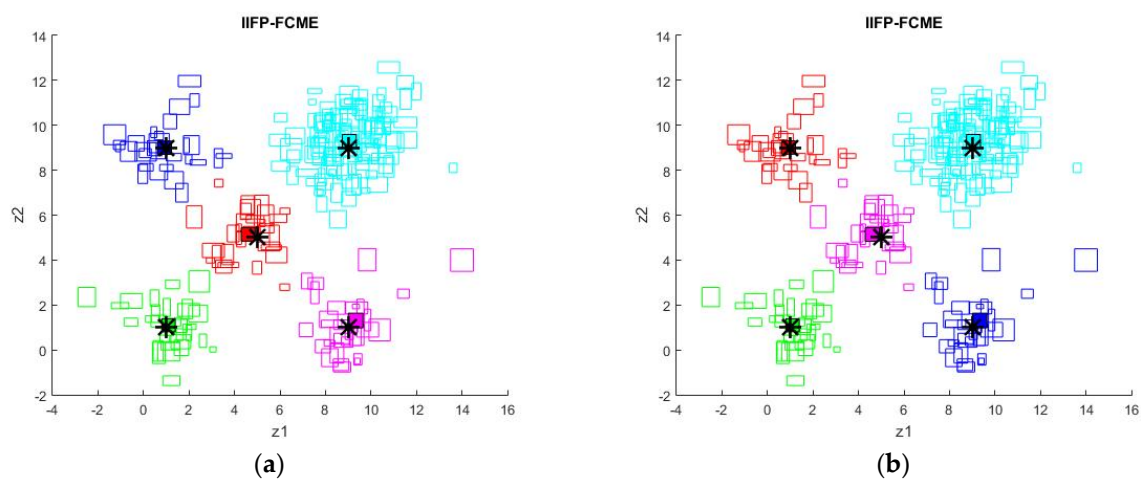


Figure 11. The results of the IIFPFCME parameter with (a) $\eta = 0.01$ and (b) $\eta = 0.1$ for IDS2.

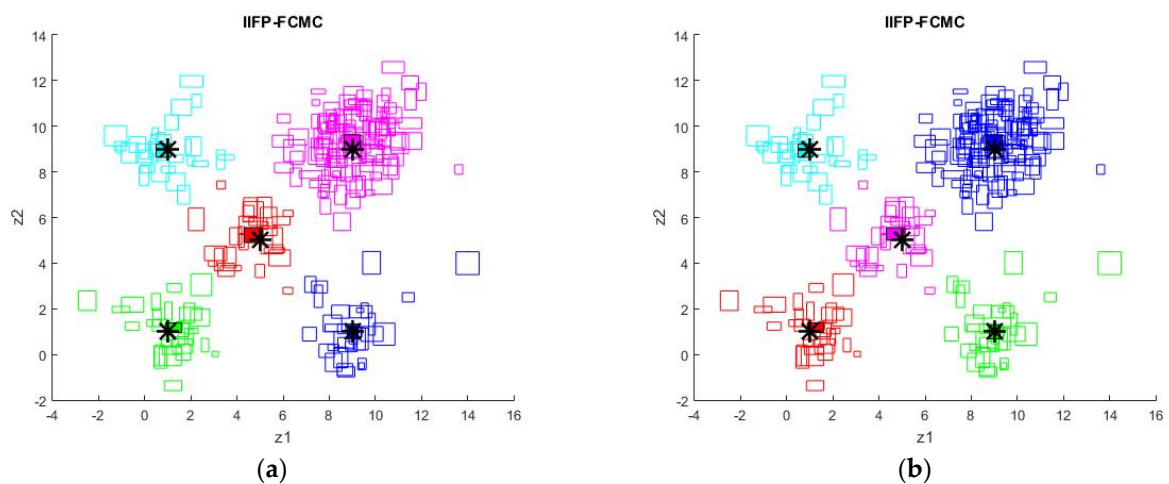


Figure 12. The results of the IIFPFCMC parameter with (a) $\eta = 0.01$ and (b) $\eta = 0.1$ for IDS2.

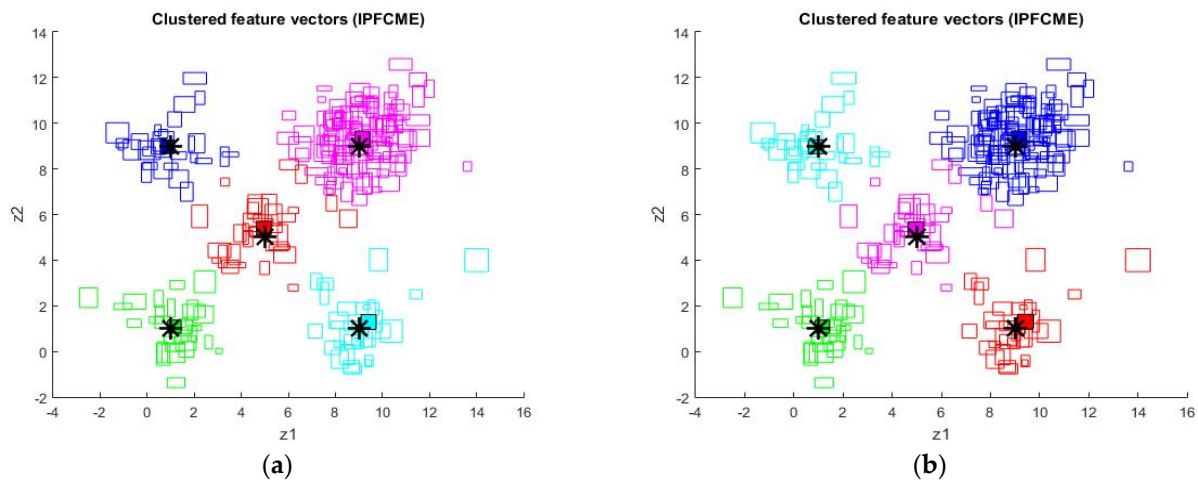


Figure 13. The results of the IPFCME parameter with (a) $m = 1.5$, $n = 2$, $a = 20$, $b = 50$ and (b) $m = 1.5$, $n = 2$, $a = 50$, $b = 50$ for IDS2.

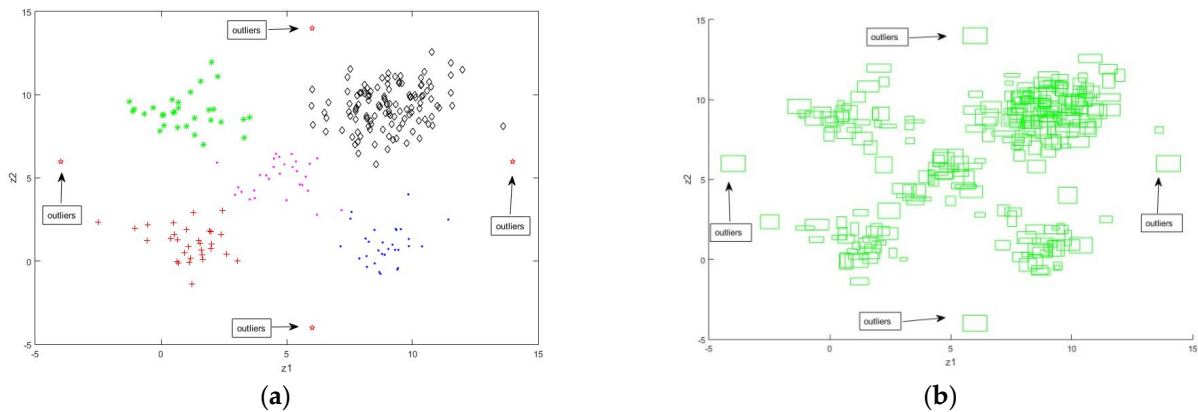


Figure 14. (a) The third dataset (DS3) and (b) interval datasets of DS3 (IDS3).

The results of IFCME and IFCMC with parameter $m = 2$ for IDS3 are shown in Figure 15. The results of IIFPFCME and IIFPFCMC with parameters $\eta = 0.01$ and $\eta = 0.1$ for IDS3 are shown in Figures 16 and 17, respectively. The results of the comparison are shown in Table 3. From Table 3, the proposed IIFPFCME and IIFPFCMC on performance and average times are all better than IFCME, IFCMC, and IPFCME. From Table 3, the proposed IIFPFCMC is better than the proposed IIFPFCME under a symbolic set with more outliers and different positions.

The fourth dataset (DS4) with 249 points consists of five clusters whose centers are (1, 1), (1, 9), (9, 1), (5, 5), and (9, 9) and diamond shapes on each group. For the first four clusters, the number of data points per cluster was 41 data points. For the last cluster centered at (9, 9), there were 85 data points (see Figure 18a); that is, DS4 used large and small groups and diamond shapes on each group to demonstrate the advantages of the proposed method. In order to build the interval datasets, this was defined as $([z1 - \gamma_1/2, z1 + \gamma_1/2], [z2 - \gamma_2/2, z2 + \gamma_2/2])$, from which γ_1 and γ_2 were drawn randomly from [0.2, 1] (see Figure 18b).

The results of IFCME and IFCMC with parameter $m = 2$ for IDS4 are shown Figure 19. The results of IIFPFCME and IIFPFCMC with parameters $\eta = 0.01$ and $\eta = 0.1$ for IDS4 are shown Figures 20 and 21, respectively. Table 4 shows the results of the comparison for IDS4 with different approaches, namely IFCME, IIFPFCME under different η , IFCM, and IIFPFCMC under different η . From Table 4, the proposed IIFPFCME and IIFPFCMC on performance and average times are all better than IFCME and IFCMC. From Table 4, the

proposed IIFPFCMC is better than the proposed IIFPFCME under a symbolic set of large and small groups and diamond shapes on each group on the RMSE.

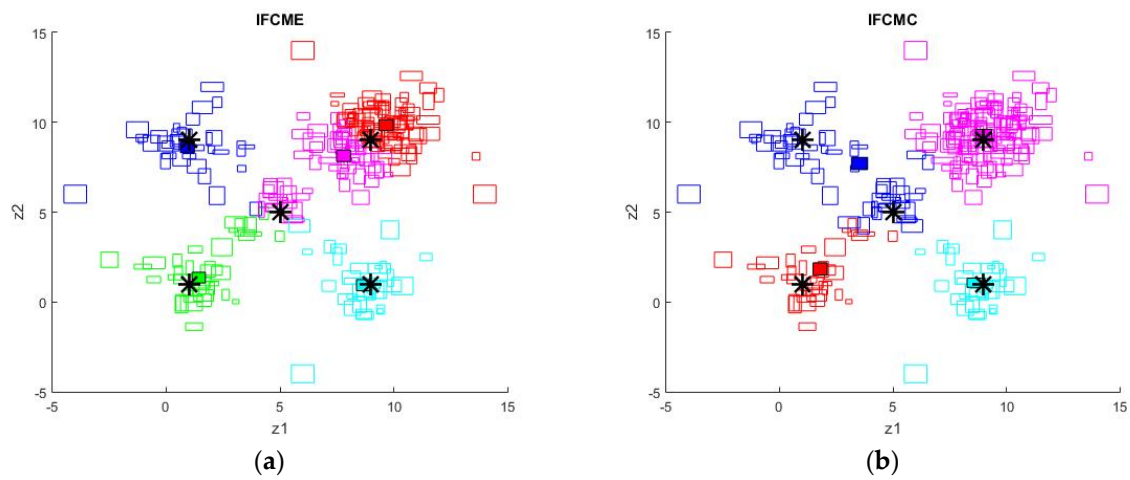


Figure 15. The results of the (a) IFCME and (b) IFCMC parameters with $m = 2$ for IDS3.

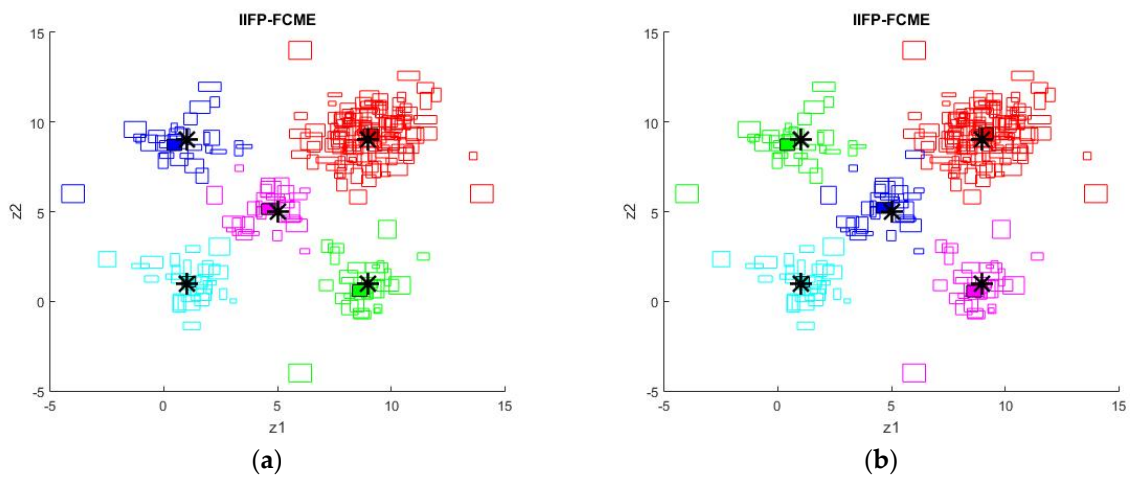


Figure 16. The results of the IIFPFCME parameter with (a) $\eta = 0.01$ and (b) $\eta = 0.1$ for IDS3.

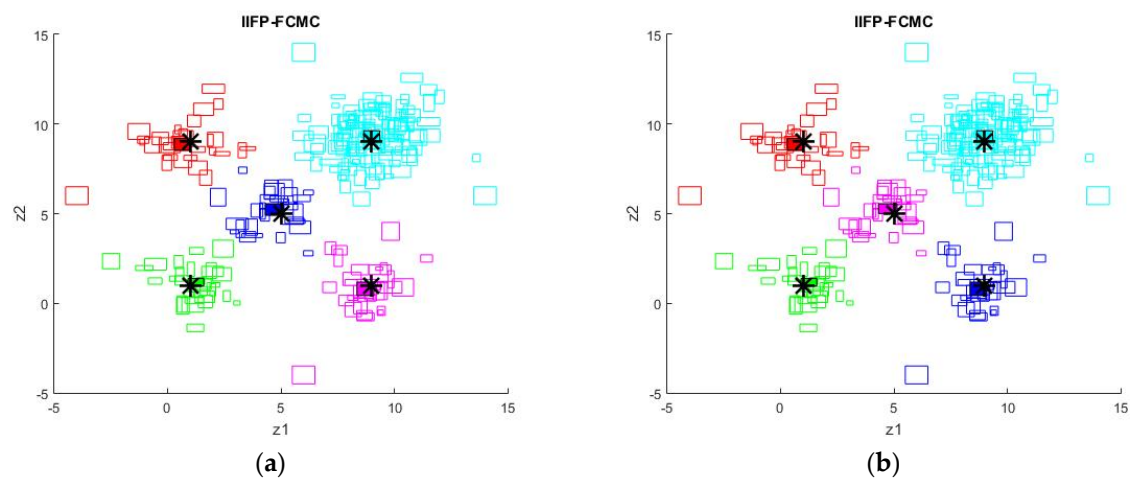


Figure 17. The results of the IIFPFCMC parameter with (a) $\eta = 0.01$ and (b) $\eta = 0.1$ for IDS3.

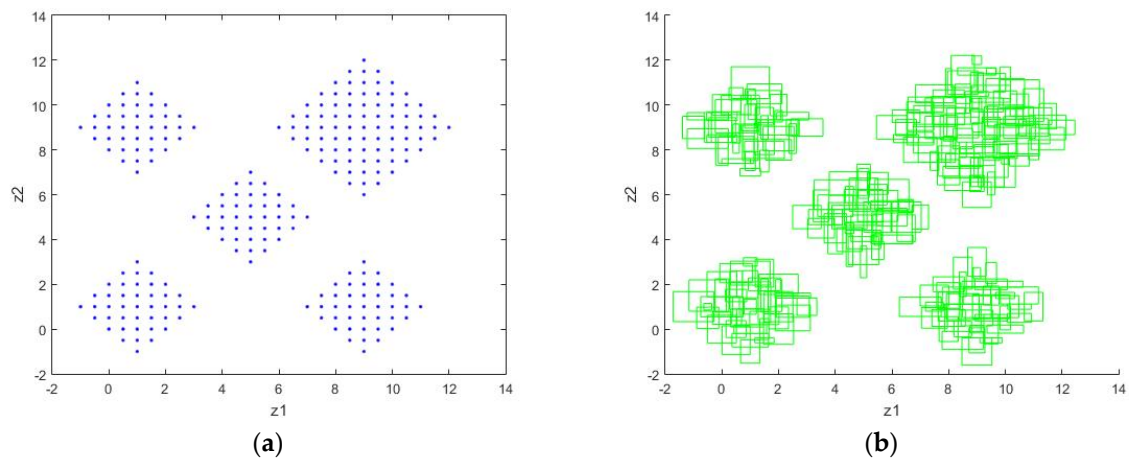


Figure 18. (a) The fourth dataset (DS4) and (b) interval datasets of DS4 (IDS4).

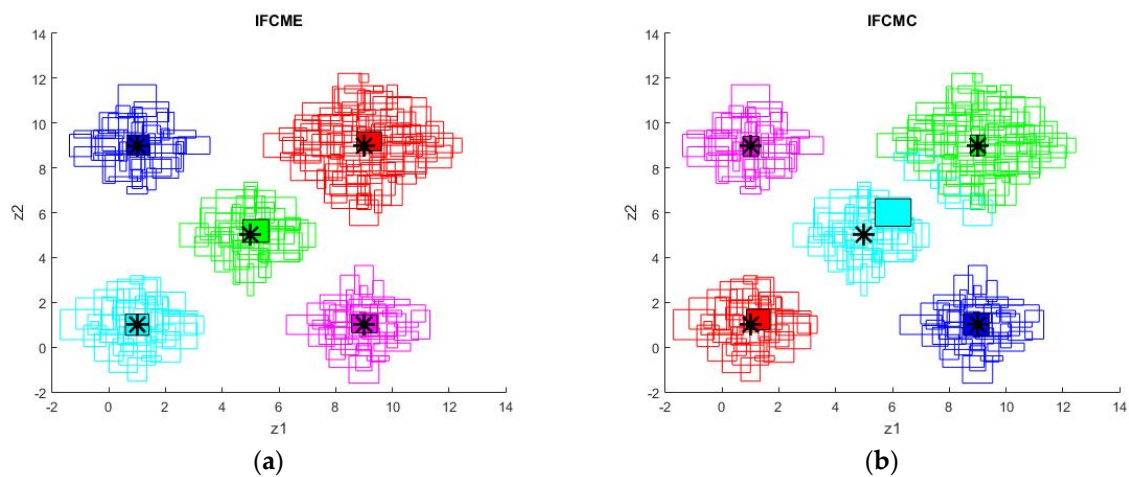


Figure 19. The results of the (a) IFCME and (b) IFCMC parameters with $m = 2$ for IDS4.

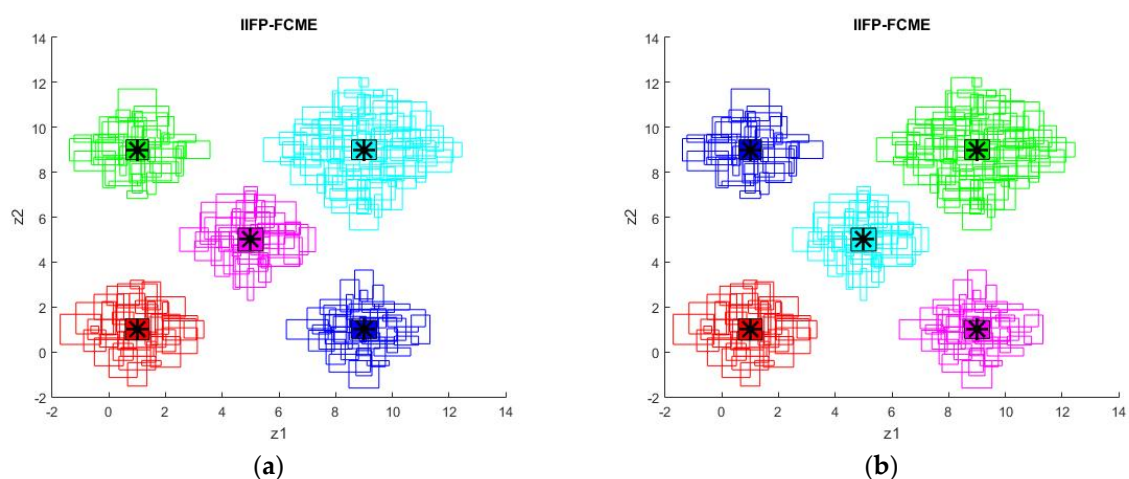


Figure 20. The results of the IIFPCME parameter with (a) $\eta = 0.01$ and (b) $\eta = 0.1$ for IDS4.

The fifth dataset (DS5) with 252 points were constructed using DS4 and another one-point (14, 4) outlier, as shown as Figure 22a. On the other hand, the interval datasets of DS5 (IDS5) are shown in Figure 22b; that is, IDS5 used large and small groups, outliers, and diamond shapes on each group to demonstrate the advantages of the proposed method.

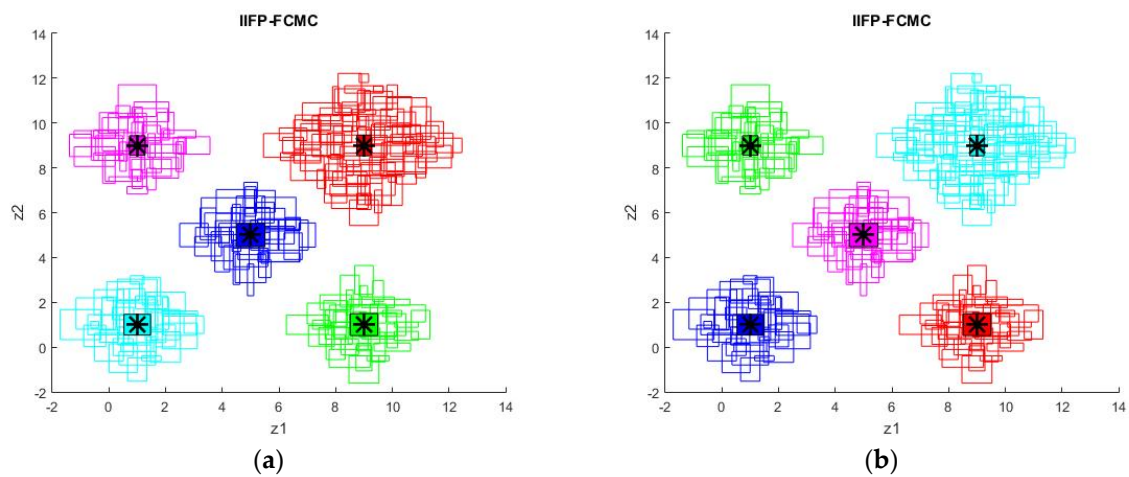


Figure 21. The results of the IIFPFCMC parameter with (a) $\eta = 0.01$ and (b) $\eta = 0.1$ for IDS4.

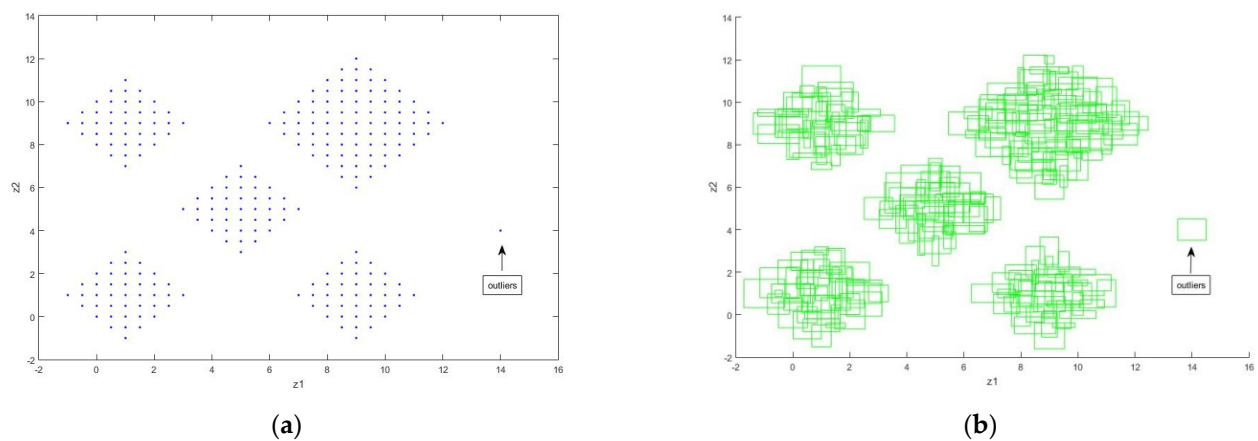


Figure 22. (a) The fifth dataset (DS5) and (b) interval datasets of DS5 (IDS5).

The results of IFCME and IFCMC with parameter $m = 2$ for IDS5 are shown in Figure 23. The results of IIFPFCME and IIFPFCMC with parameters $\eta = 0.01$ and $\eta = 0.1$ for IDS5 are shown in Figures 24 and 25, respectively. Table 5 shows the results of the comparison for IDS5 with different approaches, namely IFCME, IIFPFCME under different η , IFCMC, and IIFPFCMC under different η . From Table 5, the proposed IIFPFCME and IIFPFCMC on performance and average times are all better than IFCME and IFCMC. From Table 5, the proposed IIFPFCMC is better than the proposed IIFPFCME under a symbolic set of large and small groups, outliers, and diamond shapes on each group on the RMSE.

The sixth dataset (DS6) with 261 points were constructed using DS4 and added four points, namely $(14, 6)$, $(6, 14)$, $(6, -4)$, and $(-4, 6)$, as shown as Figure 26a. On the other hand, the interval datasets of DS6 (IDS6) are shown in Figure 26b. That is, IDS6 used large and small groups, more outliers, and diamond shapes on each group to demonstrate the advantages of the proposed method.

The results of IFCME and IFCMC with parameter $m = 2$ for IDS6 are shown in Figure 27. The results of IIFPFCME and IIFPFCMC with parameters $\eta = 0.01$ and $\eta = 0.1$ for IDS6 are shown in Figures 28 and 29, respectively. Table 6 shows the results of the comparison for IDS6 with different approaches, namely IFCME, IIFPFCME under different η , IFCMC, and IIFPFCMC under different η . From Table 6, the proposed IIFPFCME and IIFPFCMC on performance and average times are all better than IFCME and IFCMC. From Table 6, the proposed IIFPFCMC is better than the proposed IIFPFCME under a symbolic set of large and small groups, more outliers, and diamond shapes on each group on the RMSE.

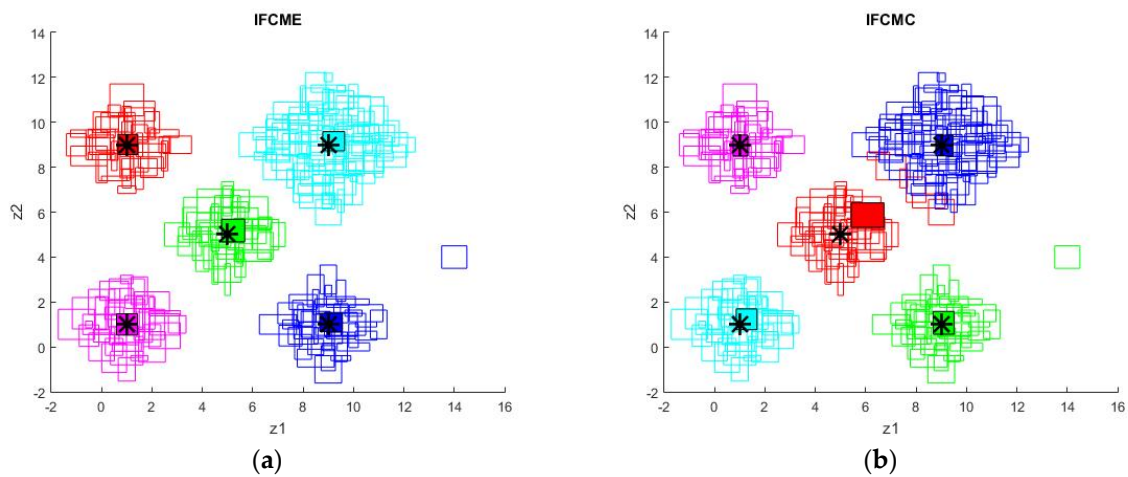


Figure 23. The results of the (a) IFCME and (b) IFCMC parameters with $m = 2$ for IDS5.

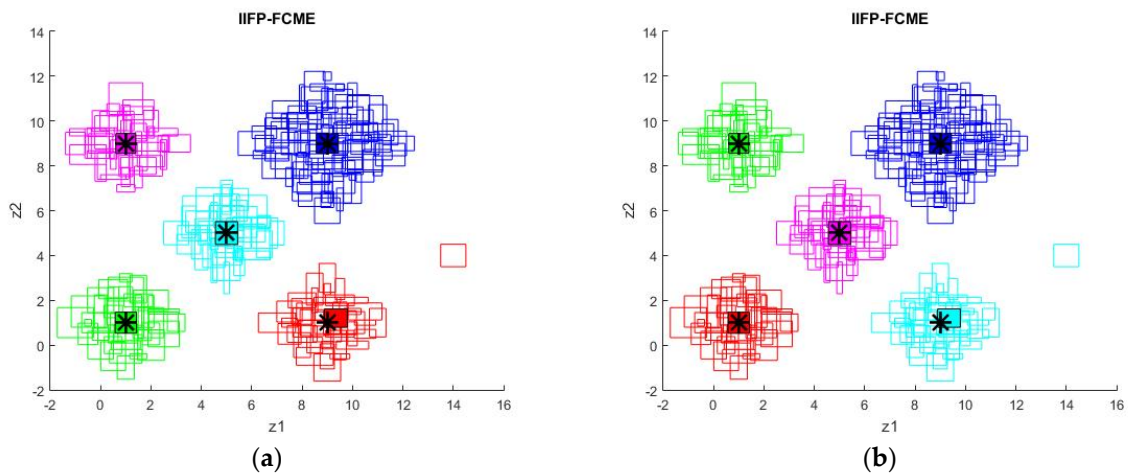


Figure 24. The results of the IIFPFCME parameter with (a) $\eta = 0.01$ and (b) $\eta = 0.1$ for IDS5.

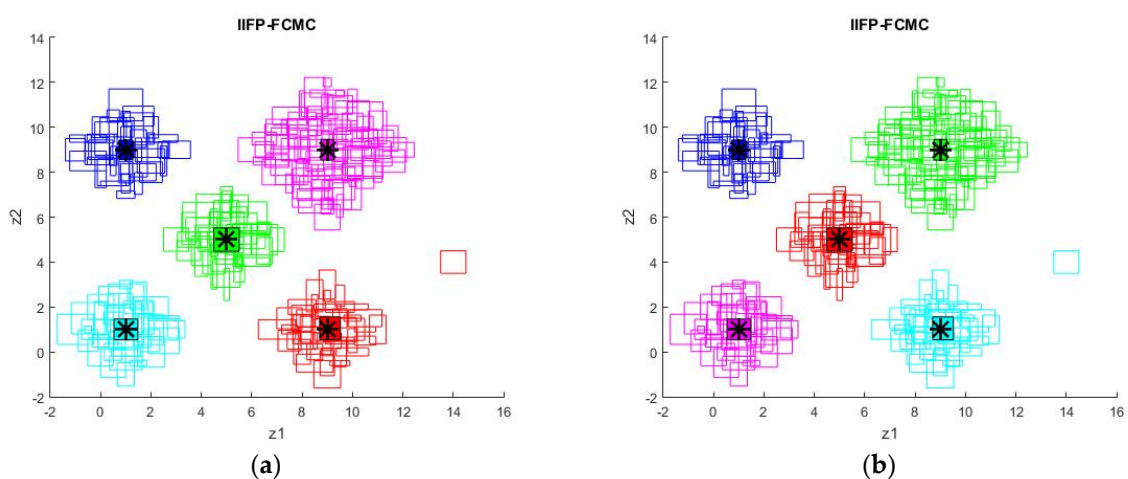


Figure 25. The results of the IIFPFCMC parameter with (a) $\eta = 0.01$ and (b) $\eta = 0.1$ for IDS5.

We presented two real datasets concerning exchange rates for validation of the proposed methods. The seventh interval dataset (IDS7) represents weekly low- and high-price correlation data for EUR/USD on z_2 and TW/USD on z_1 . The eighth interval dataset (IDS8) represents weekly low- and high-price correlation data for GBP/USD on z_2 and TW/USD on z_1 . For the currency exchange rate dataset, data were collected during a 5-year period

(2011–2015). For a detailed description of this dataset on currency exchange rates and the results of IFCME, please refer to the study published by the authors of [35].

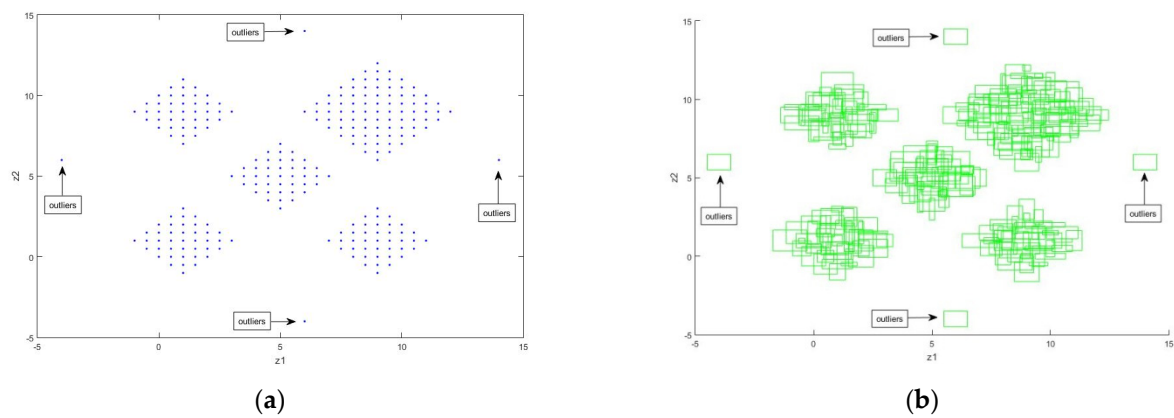


Figure 26. (a) The sixth dataset (DS6) and (b) interval datasets of DS6 (IDS6).

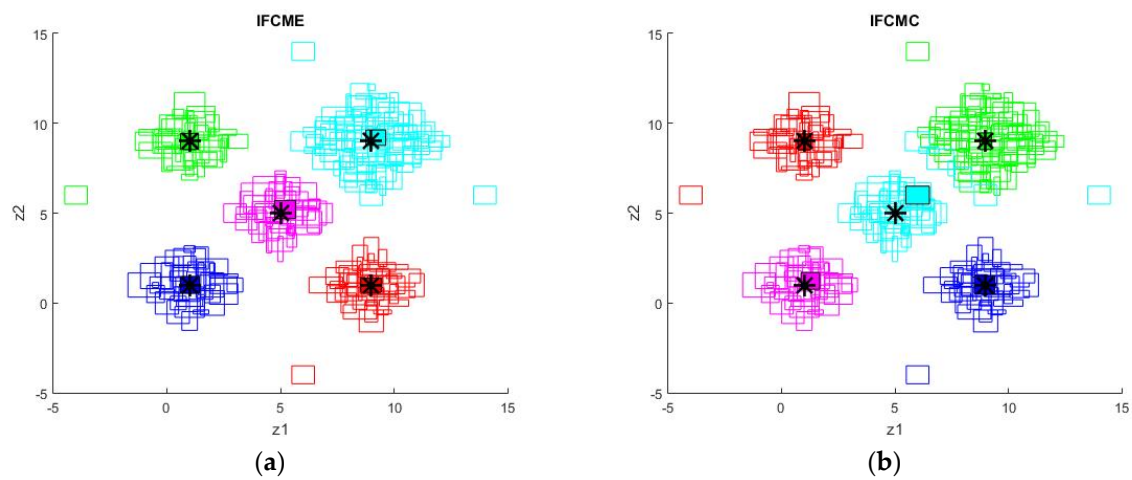


Figure 27. The results of the (a) IFCME and (b) IFCMC parameters with $m = 2$ for IDS6.

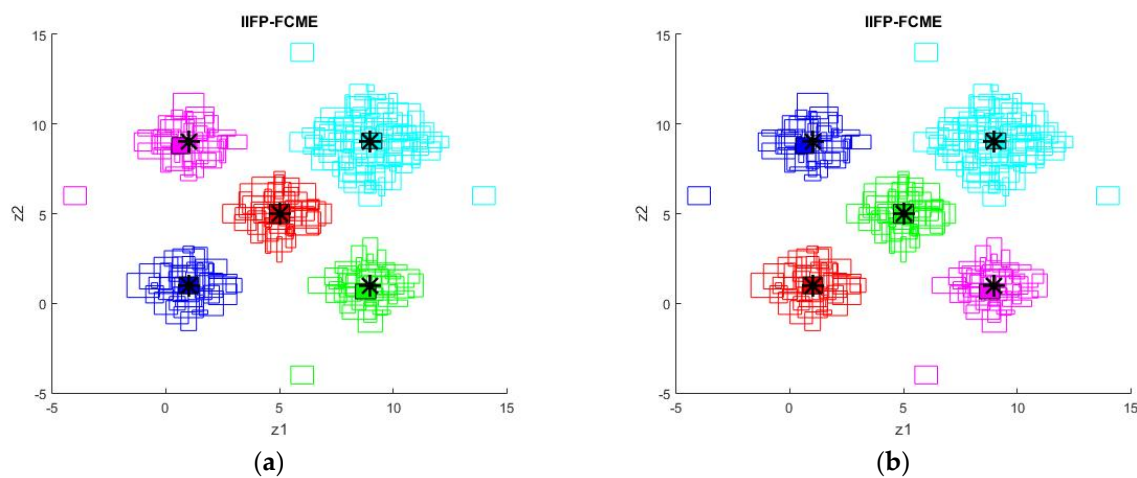


Figure 28. The results of the IIFPFCME parameter with (a) $\eta = 0.01$ and (b) $\eta = 0.1$ for IDS6.

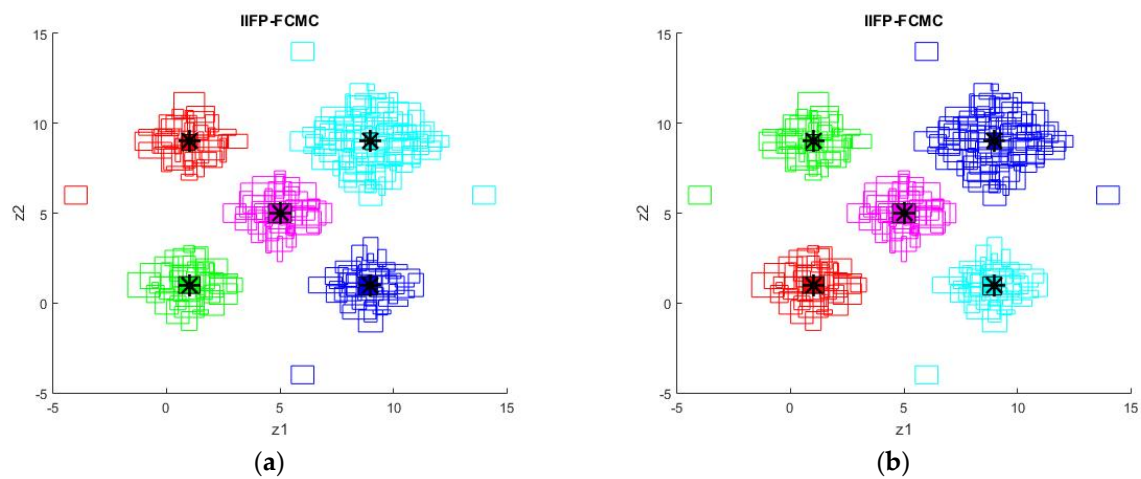


Figure 29. The results of the IIFPFCMC parameter with (a) $\eta = 0.01$ and (b) $\eta = 0.1$ for IDS6.

The results of IFCMC with parameter $m = 2$ for IDS7 are shown in Figure 30a. The results of IIFPFCMC and IIFPFCME with parameters $\eta = 1$, $\eta = 0.1$, and $\eta = 0.01$ for IDS7 are shown in Figures 30b–d and 31 respectively. Table 7 shows the results of the comparison for IDS7 with different approaches: (a) using IFCME and IIFPFCME under different η ; and (b) using IFCMC and IIFPFCMC under different η . From Table 7, the proposed IIFPFCME and IIFPFCMC appear better than the IFCME and IFCMC, respectively. Note that due to it being a real dataset, there is no information regarding the cluster center; the results only show the comparison on average times on ten (sec.) and average number of ten iterations (l) and no RMSE in the table.

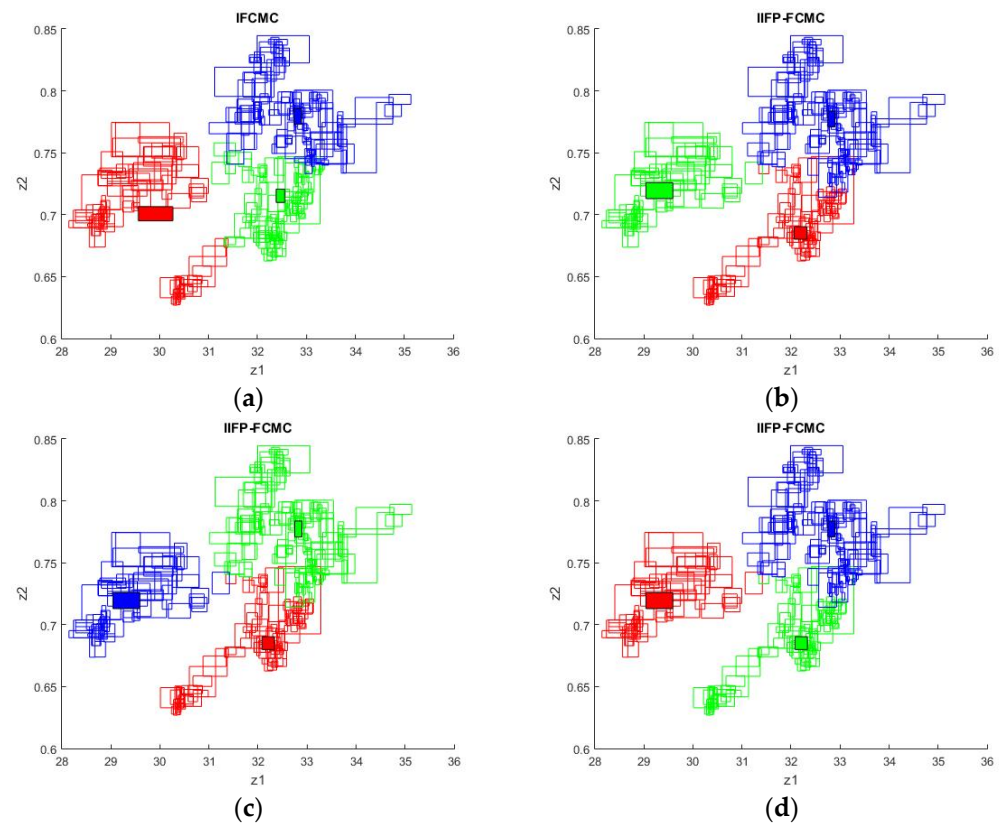


Figure 30. The results of IFCMC. (a) The results of the IIFPFCME parameter with (b) $\eta = 1$, (c) $\eta = 0.1$, and (d) $\eta = 0.01$ for IDS7.

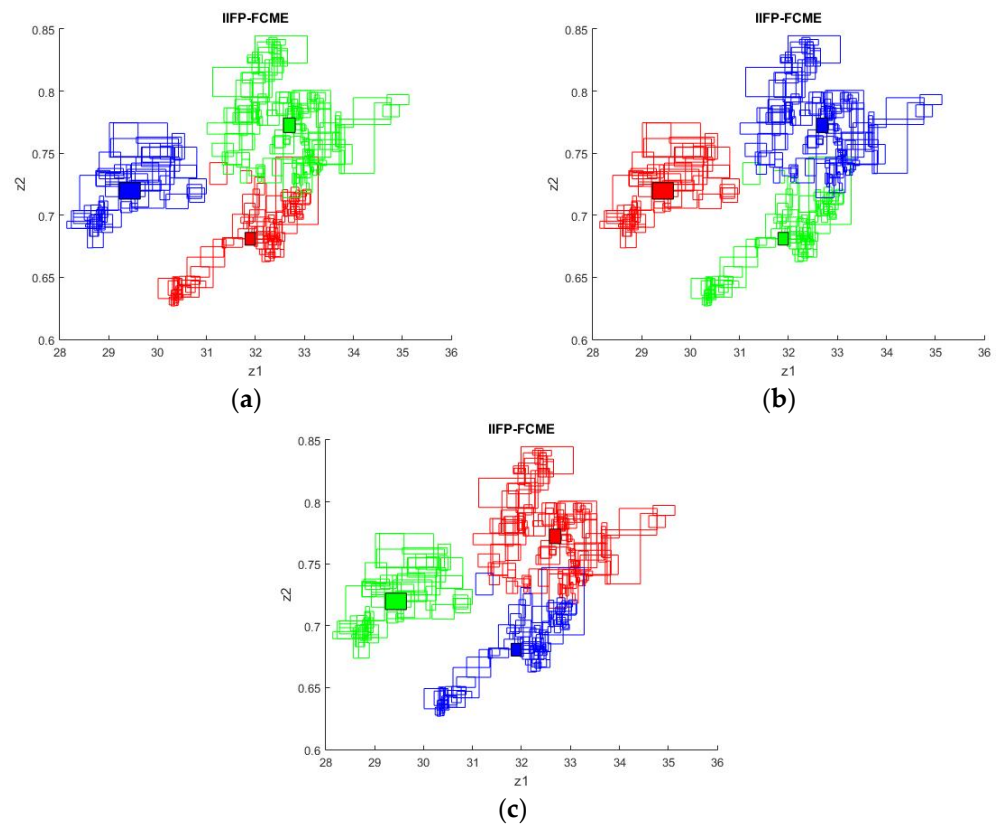


Figure 31. The results of the IIFPFCME parameter with (a) $\eta = 1$, (b) $\eta = 0.1$, and (c) $\eta = 0.01$ for IDS7.

Table 7. The results of the comparison for IDS7 with different approaches: (a) using IFCME and IIFPFCME under different η ; and (b) using IFCMC and IIFPFCMC under different η .

(a)				
Compared Index/Method	IFCME ($m = 2$)	IIFPFCME ($\eta = 0.01$)	IIFPFCME ($\eta = 0.1$)	IIFPFCME ($\eta = 1$)
Average times on ten (sec.)	0.011177	0.004754	0.005224	0.005234
Average number of ten iterations (l)	39.6	13	16.4	16.5
(b)				
Compared Index/Method	IFCMC ($m = 2$)	IIFPFCMC ($\eta = 0.01$)	IIFPFCMC ($\eta = 0.1$)	IIFPFCMC ($\eta = 1$)
Average times on ten (sec.)	0.008484	0.008062	0.009802	0.009563
Average number of ten iterations (l)	13.9	10.3	11.4	10.4

The results of IFCMC with parameter $m = 2$ for IDS8 are shown in Figure 32a. The results of IIFPFCMC and IIFPFCME with parameters $\eta = 1$, $\eta = 0.1$, and $\eta = 0.01$ for IDS7 are shown in Figures 32b–d and 33, respectively. Table 8 shows the results of the comparison for IDS7 with different approaches: (a) using IFCME and IIFPFCME under different η ; and (b) using IFCMC and IIFPFCMC under different η . From Table 8, the proposed IIFPFCME and IIFPFCMC appear better than the IFCME and IFCMC, respectively.

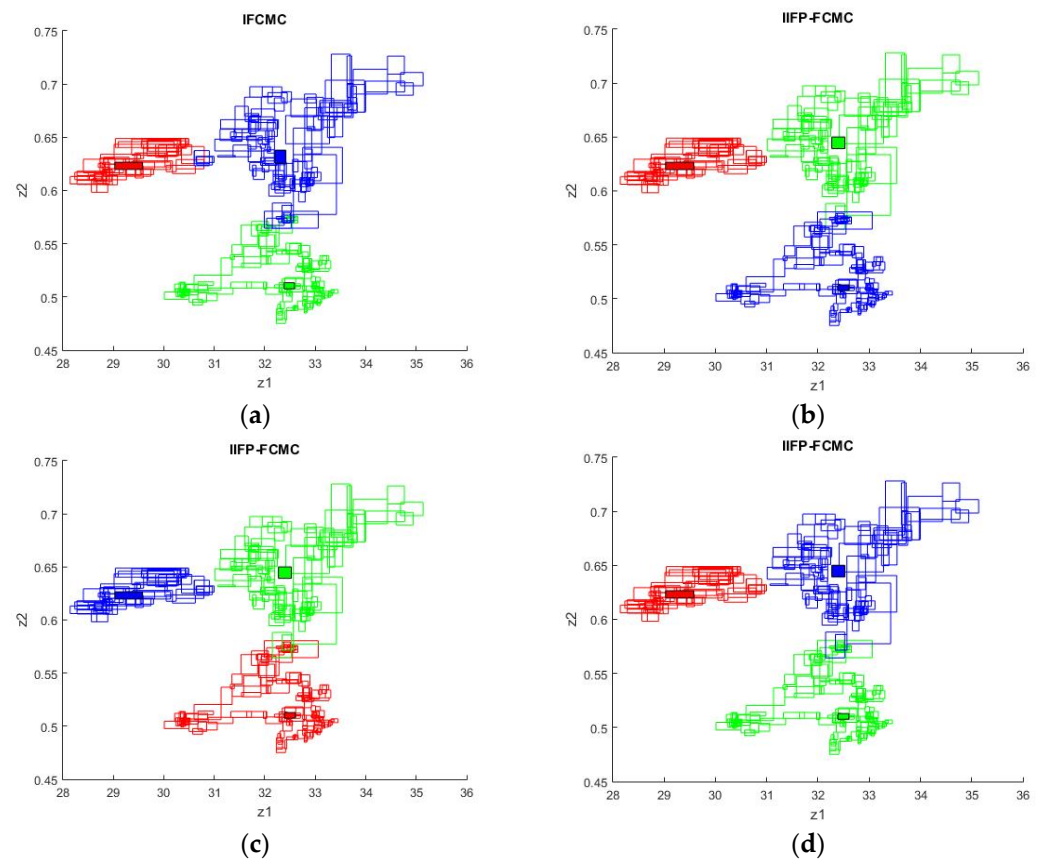


Figure 32. The results of IFCMC. (a) The results of the IIFPFCME parameter with (b) $\eta = 1$, (c) $\eta = 0.1$, and (d) $\eta = 0.01$ for IDS8.

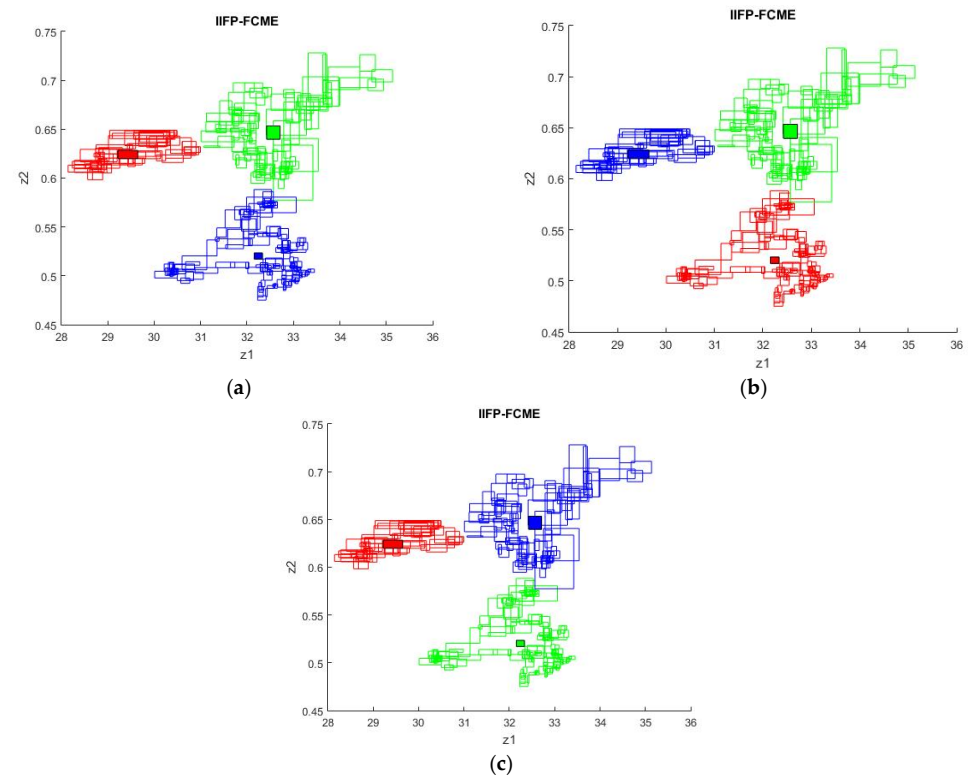


Figure 33. The results of the IIFPFCME parameter with (a) $\eta = 1$, (b) $\eta = 0.1$, and (c) $\eta = 0.01$ for IDS8.

Table 8. The results of the comparison for IDS8 with different approaches: (a) using IFCME and IIFPFCME under different η ; and (b) using IFCMC and IIFPFCMC under different η .

(a)				
Compared Index/Method	IFCME ($m = 2$)	IIFPFCME ($\eta = 0.01$)	IIFPFCME ($\eta = 0.1$)	IIFPFCME ($\eta = 1$)
Average times on ten (sec.)	0.007042	0.003014	0.003111	0.003673
Average number of ten iterations (l)	23.2	8.2	8.1	9.7
(b)				
Compared Index/Method	IFCMC ($m = 2$)	IIFPFCMC ($\eta = 0.01$)	IIFPFCMC ($\eta = 0.1$)	IIFPFCMC ($\eta = 1$)
Average times on ten (sec.)	0.007925	0.007317	0.006542	0.007304
Average number of ten iterations (l)	13.5	8.8	7.8	8

In order to assess the efficiency of the proposed methods, the silhouette index (SI) was used as a clustering validation index. The SI is a metric used to calculate the goodness of a clustering technique. Its value provides a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The SI is particularly useful under scenarios where the true labels of the data are not known [38,39]. It provides a way to compare the effectiveness of different clustering methods on a given dataset:

$$S_{SI} = \frac{b_{SI} - a_{SI}}{\max(a_{SI}, b_{SI})} \quad (30)$$

where a_{SI} is the mean distance to the points in the same cluster, and b_{SI} is the mean distance to the points in the nearest other cluster. Regarding the interpretation of the SI, (a) the silhouette coefficient ranges from -1 to $+1$; (b) a high value indicates that the object is well matched to its own cluster and poorly matched to their neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate; (c) if many points have a low or negative value, then the clustering configuration may have too many or too few clusters. Due to IDS7 and IDS8 being real datasets, there is no information regarding their cluster centers. Hence, the SI was used as a clustering validation index to assess the efficiency of the proposed methods. The compared results used the SI of IDS7 with IFCME and the proposed IIFPFCME with different η in Table 9. The compared results used the SI of IDS8 with IFCME and the proposed IIFPFCME with different η in Table 10. Tables 9 and 10 show that the proposed methods have good results in terms of efficiency with the SI.

Table 9. The compared results used the SIs of IDS7–8 with IFCME and the proposed IIFPFCME with different η .

Silhouette Index/Method	IFCME ($m = 2$)	IIFPFCME ($\eta = 0.01$)	IIFPFCME ($\eta = 0.1$)	IIFPFCME ($\eta = 1$)
SI of IDS7	0.613169	0.670552	0.669509	0.669509
SI of IDS8	0.678284	0.764616	0.764616	0.764616

Table 10. The compared results used the SIs of IDS7–8 with IFCMC and the proposed IIFPFCMC with different η .

Silhouette Index/Method	IFCMC ($m = 2$)	IIFPFCMC ($\eta = 0.01$)	IIFPFCMC ($\eta = 0.1$)	IIFPFCMC ($\eta = 1$)
SI of IDS7	0.612450	0.655276	0.655276	0.655276
SI of IDS8	0.733026	0.760424	0.760424	0.760424

In the above tests, we used three types of data points and added outlier data to these three types for testing. The algorithm of IIFPFCMC exhibits better clustering results in the diamond-shaped and square-shaped distribution data points, and the outlier data were added to these three types of data. The performance and average of ten iterations of the IIFPFCMC algorithm was also found to be better. Although the number of iterations of the proposed IIFPFCMC was reduced, it takes more computing time to obtain the median. The results displayed in all tables were based on converging to the best group center point to count the time and the number of iterations. Hence, the proposed IIFPFCM clustering algorithm with the city block distance measure was deemed to be suitable for dealing with SIDA with outliers and exhibits a better performance and average of iterations than IFCME, IFCMC, and IPFCME. Additionally, the results of the real datasets also indicate that the proposed methods have good results in terms of efficiency with the SI.

6. Conclusions

This study presented the IIFPFCM clustering algorithm, focusing on achieving fast convergence through independent combinations with both Euclidean distance and city block distance. Both proposed methods exhibited faster convergence speeds compared to the traditional IFCM and IPFCM clustering methods in SIDA. Furthermore, there was an issue of partitioning into large and small groups in the context of SID. The proposed methods also yielded superior performance results compared to the traditional IFCM clustering method for addressing this challenge. Additionally, traditional IFCM clustering is susceptible to outliers. This study also introduced the IIFPFCM algorithm, which tackles outliers by employing interval distance measurements. In other words, we enhanced the convergence speed of IFCM, reduced the computation time, and achieved a more effective handling of outliers. In our experiments, we compared three types of data points and introduced outlier data to each of these types. The results of these experiments demonstrated that IIFPFCME effectively enhances convergence speed and the partitioning of data into large and small groups. However, it is worth noting that IIFPFCMC effectively handles outlier data but is sensitive to the choice of initial values in this study. Additionally, the proposed algorithm is currently not affiliated with any commercial entities, and its development remains in the theoretical stage. Regarding the directions for further development of the presented algorithms, SIDA will be employed in many sectors, like healthcare, marketing, education, etc., since SIDA usually provides analysts with a large amount of data on behavior. Moreover, in robotics and autonomous systems, SIDA can aid in interpreting these data for navigation, obstacle avoidance, and decision-making processes. In agriculture, SIDA can aid in soil classification, crop monitoring, and yield prediction by handling the imprecision inherent in agricultural data. In the context of market research, SIDA can be employed to segment markets and understand consumer groups based on imprecise data like survey responses. In bioinformatics, SIDA can assist in grouping similar genetic profiles, which can be useful in understanding diseases and developing treatments.

Author Contributions: Conceptualization, S.-C.C. and J.-T.J.; methodology, S.-C.C. and J.-T.J.; programming, S.-C.C.; writing—original draft preparation, S.-C.C. and J.-T.J.; writing—review and editing, W.-C.C. and J.-T.J.; funding acquisition, J.-T.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Science Council Under Grant NSTC 110-2221-E-150-040.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in article.

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

Abbreviations

DS	Datasets
FCM	Fuzzy C-means
FMD	Fuzzy membership degree
IDS	Interval datasets
IFCMC	Interval fuzzy C-means with city block distance
IFCME	Interval fuzzy C-means with Euclidean distance
IFPFCM	Improved fuzzy partitions for fuzzy regression models
IIFPFCM	Interval improved fuzzy partitions fuzzy C-means
IIFPFCMC	IIFPFCM with city block distance
IIFPFCME	IIFPFCM with Euclidean distance
IPFCM	Interval possibilistic FCM
PMD	Possibilistic membership degree
RMSE	Root-mean-square error
SID	Symbolic interval data
SIDA	Symbolic interval data analysis

References

- Billard, L.; Diday, E. *Symbolic Data Analysis: Conceptual Statistics and Data Mining*; John Wiley & Sons: London, UK, 2006.
- Chuang, C.-C.; Jeng, J.-T.; Chang, S.-C. Hausdorff distance measure based interval fuzzy possibilistic c-means clustering algorithm. *Int. J. Fuzzy Syst.* **2013**, *15*, 471–479.
- He, Q.; He, Z.; Duan, S.; Zhong, Y. Multi-objective interval portfolio optimization modeling and solving for margin trading. *Swarm Evol. Comput.* **2022**, *75*, 101141. [\[CrossRef\]](#)
- Zhou, B.; Wang, X.; Zhou, J.; Jing, C. Trajectory recovery based on interval forward–backward propagation algorithm fusing multi-source information. *Electronics* **2022**, *11*, 3634. [\[CrossRef\]](#)
- Yamaka, W.; Phadkantha, R.; Maneejuk, P. A convex combination approach for artificial neural network of interval data. *Appl. Sci.* **2021**, *11*, 3997. [\[CrossRef\]](#)
- Fordellone, M.; De Benedictis, I.; Bruzzese, D.; Chiodini, P. A maximum-entropy fuzzy clustering approach for cancer detection when data are uncertain. *Appl. Sci.* **2023**, *13*, 2191. [\[CrossRef\]](#)
- Freitas, W.W.F.; Souza, R.M.C.R.; Getúlio, J.A.; Bastian, F. Exploratory spatial analysis for interval data: A new autocorrelation index with COVID-19 and rent price applications. *Expert Syst. Appl.* **2022**, *195*, 116561. [\[CrossRef\]](#)
- Chang, W.; Ji, X.; Liu, Y.; Xiao, Y.; Chen, B.; Liu, H.; Zhou, S. Analysis of university students' behavior based on a fusion k-means clustering algorithm. *Appl. Sci.* **2020**, *10*, 6566. [\[CrossRef\]](#)
- Zhang, R.-L.; Liu, X.-H. A novel hybrid high-dimensional pso clustering algorithm based on the cloud model and entropy. *Appl. Sci.* **2023**, *13*, 1246. [\[CrossRef\]](#)
- Dougherty, E.R.; Brun, M. A probabilistic theory of clustering. *Pattern Recognit. Soc.* **2004**, *37*, 917–925. [\[CrossRef\]](#)
- Volkovich, Z.; Avros, R.; Golani, M. On initialization of the expectation maximization clustering algorithm. *Glob. J. Technol. Optim.* **2011**, *2*, 1–4.
- Sun, T.; Shu, C.; Li, F.; Yu, H.; Ma, L.; Fang, Y. An efficient hierarchical clustering method for large datasets with map-reduce. In Proceedings of the 2009 International Conference on Parallel and Distributed Computing, Applications and Technologies, Boston, MA, USA, 24–26 September 2009; pp. 494–499.
- Li, M.; Deng, S.; Wang, L.; Feng, S.; Fan, J. Hierarchical clustering algorithm for categorical data using a probabilistic rough set model. *Knowl. Based Syst.* **2014**, *65*, 60–71. [\[CrossRef\]](#)
- Patel, S.; Sihmar, S.; Jatain, A. A study of hierarchical clustering algorithms. In Proceedings of the 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 11–13 March 2015; pp. 537–541.
- Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A k-means clustering algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.* **1979**, *28*, 100–108. [\[CrossRef\]](#)
- Park, H.S.; Jun, C.H. A simple and fast algorithm for K-medoids clustering. *Expert Syst. Appl.* **2009**, *36*, 3336–3341. [\[CrossRef\]](#)
- Fahad, A.; Alshatri, N.; Tari, Z.; Alamri, A.; Khalil, I.; Zomaya, A.Y.; Foufou, S.; Bouras, A. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Trans. Emerg. Top. Comput.* **2014**, *2*, 267–279. [\[CrossRef\]](#)
- Bezdek, J.C.; Ehrlich, R.; Full, W. FCM: The fuzzy c-means clustering algorithm. *Comput. Geosci.* **1984**, *10*, 191–203. [\[CrossRef\]](#)
- Mújica-Vargas, D.; Kinani, J.M.V.; Rubio, J.D. Color-based image segmentation by means of a robust intuitionistic fuzzy c-means algorithm. *Int. J. Fuzzy Syst.* **2020**, *22*, 901–916. [\[CrossRef\]](#)
- Gao, Y.; Li, H.; Li, J.; Cao, C.; Pan, J. Patch-based fuzzy local weighted c-means clustering algorithm with correntropy induced metric for noise image segmentation. *Int. J. Fuzzy Syst.* **2023**, *25*, 1991–2006. [\[CrossRef\]](#)
- Hussain, I.; Sinaga, K.P.; Yang, M.-S. Unsupervised multiview fuzzy c-means clustering algorithm. *Electronics* **2023**, *12*, 4467. [\[CrossRef\]](#)
- Shi, Y. Application of FCM clustering algorithm in digital library management system. *Electronics* **2022**, *11*, 3916. [\[CrossRef\]](#)

23. Tang, Y.; Chen, R.; Xia, B. VSFCM: A novel viewpoint-driven subspace fuzzy c-means algorithm. *Appl. Sci.* **2023**, *13*, 6342. [\[CrossRef\]](#)
24. Wang, Y.; Qin, Q.; Zhou, J.; Chen, Y.; Han, S.; Wang, L.; Du, T.; Ji, K.; Zhao, Y.O.; Zhang, K. Guided filter-based fuzzy clustering for general data analysis. *Int. J. Fuzzy Syst.* **2023**, *25*, 2036–2051. [\[CrossRef\]](#)
25. Sousa, Á.; Silva, O.; Bacelar-Nicolau, L.; Cabral, J.; Bacelar-Nicolau, H. Comparison between two algorithms for computing the weighted generalized affinity coefficient in the case of interval data. *Stats* **2023**, *6*, 1082–1094. [\[CrossRef\]](#)
26. Roh, S.B.; Oh, S.K.; Pedrycz, W.; Wang, Z.; Fu, Z.; Seo, K. Design of iterative fuzzy radial basis function neural networks based on iterative weighted fuzzy c-means clustering and weighted LSE estimation. *IEEE Trans. Fuzzy Syst.* **2022**, *30*, 4273–4285. [\[CrossRef\]](#)
27. Huang, Y.P.; Bhalla, K.; Chu, H.C.; Lin, Y.C.; Kuo, H.C.; Chu, W.J.; Lee, J.H. Wavelet k-means clustering and fuzzy-based method for segmenting MRI images depicting Parkinson's disease. *Int. J. Fuzzy Syst.* **2021**, *23*, 1600–1612. [\[CrossRef\]](#)
28. Elsheikh, S.; Fish, A.; Zhou, D. Exploiting spatial information to enhance DTI segmentations via spatial fuzzy c-means with covariance matrix data and non-euclidean metrics. *Appl. Sci.* **2021**, *11*, 7003. [\[CrossRef\]](#)
29. Höppner, F.; Klawonn, F. Improved fuzzy partitions for fuzzy regression models. *Int. J. Approx. Reason* **2003**, *32*, 85–102. [\[CrossRef\]](#)
30. Hazarika, I.; Mahanta, A.K. A New Semimetric for Interval Data. *Int. J. Recent Technol. Eng.* **2019**, *8*, 3278–3285. [\[CrossRef\]](#)
31. De Souza, R.M.C.R.; De Carvalho, F.A.T. Clustering of interval data based on city-block distances. *Pattern Recognit. Lett.* **2004**, *25*, 353–365. [\[CrossRef\]](#)
32. De Carvalho, F.D.A.T.; Brito, P.; Bock, H.-H. Dynamic clustering for interval data based on L_2 distance. *Comput. Statist.* **2006**, *21*, 231–250. [\[CrossRef\]](#)
33. Peng, W.; Li, T. Interval Data Clustering with Applications. In Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence, Arlington, VA, USA, 13–15 November 2006; pp. 355–362.
34. De Carvalho, F.D.A.T. Fuzzy c-means clustering methods for symbolic interval data. *Pattern Recognit. Lett.* **2007**, *28*, 423–437. [\[CrossRef\]](#)
35. Jeng, J.-T.; Chen, C.-M.; Chang, S.-C.; Chuang, C.-C. IPFCM clustering algorithm under Euclidean and Hausdorff distance measure for symbolic interval data. *Int. J. Fuzzy Syst.* **2019**, *21*, 2102–2119. [\[CrossRef\]](#)
36. Chen, C.-M.; Chang, S.-C.; Chuang, C.-C.; Jeng, J.-T. Rough IPFCM clustering algorithm and its application on smart phones with Euclidean distance. *Appl. Sci.* **2022**, *12*, 5195. [\[CrossRef\]](#)
37. Kato, J.; Okada, K. Simplification and shift in cognition of political difference: Applying the geometric modeling to the analysis of semantic similarity judgment. *PLoS ONE* **2011**, *6*, e20693. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [\[CrossRef\]](#)
39. Shahapure, K.R.; Nicholas, C. Cluster Quality Analysis Using Silhouette Score. In Proceedings of the 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), Sydney, NSW, Australia, 6–9 October 2020; pp. 747–748.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.