Computer Systems Science & Engineering
DOI: 10.32604/csse.2023.035692
Article

Picture-Neutrosophic Trusted Safe Semi-Supervised Fuzzy Clustering for Noisy Data

Pham Huy Thong[1,2,3], Florentin Smarandache[4], Phung The Huan[5], Tran Manh Tuan[6], Tran Thi Ngan[6,*], Vu Duc Thai[5], Nguyen Long Giang[2] and Le Hoang Son[3]

[1]Graduate University of Science and Technology, Vietnam Academy of Science and Technology, Hanoi, 100000, Vietnam
[2]Institute of Information Technology, Vietnam Academy of Science and Technology, Hanoi, 100000, Vietnam
[3]VNU Information Technology Institute, Vietnam National University, Hanoi, 100000, Vietnam
[4]Department of Mathematics, University of New Mexico, Gallup, 87301, New Mexico, USA
[5]University of Information and Communication Technology, Thai Nguyen University, Thai Nguyen, 250000, Vietnam
[6]Faculty of Computer Science and Engineering, Thuyloi University, Hanoi, 100000, Vietnam
*Corresponding Author: Tran Thi Ngan. Email: ngantt@tlu.edu.vn
Received: 31 August 2022; Accepted: 14 December 2022

Abstract: Clustering is a crucial method for deciphering data structure and producing new information. Due to its significance in revealing fundamental connections between the human brain and events, it is essential to utilize clustering for cognitive research. Dealing with noisy data caused by inaccurate synthesis from several sources or misleading data production processes is one of the most intriguing clustering difficulties. Noisy data can lead to incorrect object recognition and inference. This research aims to innovate a novel clustering approach, named Picture-Neutrosophic Trusted Safe Semi-Supervised Fuzzy Clustering (PNTS3FCM), to solve the clustering problem with noisy data using neutral and refusal degrees in the definition of Picture Fuzzy Set (PFS) and Neutrosophic Set (NS). Our contribution is to propose a new optimization model with four essential components: clustering, outlier removal, safe semi-supervised fuzzy clustering and partitioning with labeled and unlabeled data. The effectiveness and flexibility of the proposed technique are estimated and compared with the state-of-art methods, standard Picture fuzzy clustering (FC-PFS) and Confidence-weighted safe semi-supervised clustering (CS3FCM) on benchmark UCI datasets. The experimental results show that our method is better at least 10/15 datasets than the compared methods in terms of clustering quality and computational time.

Keywords: Safe semi-supervised fuzzy clustering; picture fuzzy set; neutrosophic set; data partition with noises; fuzzy clustering

1 Introduction

The finding of underlying connections between the human brain and events has made the development of sophisticated clustering algorithms fashionable in cognitive research [1,2]. Dealing with noisy data is one of the most intriguing clustering difficulties. Incorrect data with noises that affect the quality of results could be seen in many applications, such as satellite images [3], medical image processing [4,5], control systems [6], etc.

Semi-supervised fuzzy clustering techniques were introduced with additional information provided by users [7–9] to enhance the range of applications and the quality of clusters. The differences in incorporating various supplementary information forms were demonstrated in [10] which provided a summary of the semi-supervised fuzzy clustering technique. Accordingly, object segmentation using semi-supervised fuzzy clustering is effective as long as the proper supplementary information, also known as "safe information" and clean data are supplied. However, real-world data are frequently unreliable, noisy and inaccurate. These situations require more effective clustering methods.

The safe semi-supervised fuzzy clustering approach introduced in [11–13] is the typical method to deal with safe information in semi-supervised fuzzy clustering. There are two primary phases in their strategy after the core concept. The confidence weights for labeled data are calculated in the first phase. Then, the high confidence weights are used to generate and identify centers of clusters and fuzzy element values under the labeled data in the second phase. Safe semi-supervised Fuzzy C-Means clustering (S3FCM) approach was firstly presented in [11]. By balancing semi-supervised and unsupervised clustering, this technique investigated the incorrectly classified data. A local homogeneous graph was employed in the first phase [12]. The Local Homogeneous Consistent Safe SSFCM (LHC-S3FCM) method performed effectively on datasets with a large percentage of incorrectly categorized data by utilizing this graph. The CS3FCM, an enhanced safe semi-supervised clustering model, based on confidence weights, was put out in [13]. This approach provides good results in minimizing the negative impact of incorrectly labeled samples on the clustering process, assuming each data sample has its own safe confidence weight.

To establish the safe level of each sample in the data set, Guo et al. [14] have recently suggested a safe semi-supervised clustering with a safe degree. The model provides the essential procedures to reduce the adverse effects of risk in both labeled and unlabeled samples based on the safe degree value. Despite performing better than other approaches when dealing with "safe information", safe semi-supervised fuzzy clustering algorithms can still not solve the challenge of clustering inaccurate data with noises. Noisy data division can lead to incorrect object detection and inference. Data points, isolated or at the edge of some clusters, are considered to contain noisy data. It is a must to improve safe semi-supervised fuzzy clustering algorithms for dealing with noisy data.

This research aims to develop a new clustering method to remove the noise from data and increase the performance of the clustering method. This method integrates the semi-supervised clustering method and the picture fuzzy set [15]. There are four membership degrees in the PFS [3] with Neutrosophic set [16], including the positive degree, neutral degree, negative degree and rejection degree. Noisy data typically have a high rejection rate. Additionally, the neutral degree is used to determine the data points belonging to the boundary of clusters. It is clear that PFS could be used to identify noisy data in datasets.

Based on the original Fuzzy C-Means (FCM) model, a fuzzy clustering algorithm for images (a.k.a. FC-PFS) introduced in [17] outperforms the other fuzzy clustering techniques in terms of average clustering indices such as the mean accuracy and computational time. As an extension of collaborative distributed fuzzy clustering (CDFCM) [18] on PFS, a form of FC-PFS on distributed computing known as DPFCM was demonstrated in [19]. As stated in the paper, a strategy to reduce computational time and increase clustering quality is the idea of semi-supervised clustering using distributed and cloud computing. Wu and Chen presented an adaptive picture fuzzy clustering technique based on entropy weight [20]. This approach improved accuracy, addressed noisy data in image segmentation and overcame the time-consuming limitation in existing picture fuzzy clustering algorithms. Two practical, robust picture fuzzy clustering techniques for decreasing computational time were also introduced [21,22]. Nonetheless, those fuzzy clustering algorithms struggle with managing both the "safe information" and the "noisy data" because if labeled data has noise, the clustering quality will be seriously affected.

To handle problems with enhancing "safe information" and reducing the effect of "noisy data", Picture-Neutrosophic Trusted Safe Semi-Supervised Fuzzy Clustering (PNTS3FCM) is introduced. This is a new technique to address the issue of data partition with noisy information. The PNTS3FCM approach includes picture fuzzy and neutrosophic set concepts in the semi-supervised fuzzy clustering with a safe information procedure. The research proposes a new optimization model consisting of four essential components: a clustering component, an outlier-solving component and a safe semi-supervised fuzzy clustering using labeled and unlabeled data. The first two parts employed FC-PFS and the last two are the new parts to enhance safe information and reduce noisy data. An iterative technique from the formulation is also provided to construct the cluster centers and memberships. In fact, the survey has revealed a new field of study: safe, semi-supervised clustering on the picture fuzzy set. To compare PNTS3FCM with other available methods on benchmark datasets, two similar algorithms-FC-PFS [17] and CS3FCM [13], are chosen.

The remaining paper is structured as follows: Section 2 offers the essential information underpinning our study. The proposed approach is introduced in Section 3 and the experimental results are presented in Section 4. Some conclusions are given in the last section.

2 Preliminaries

In this section, some fundamental concepts and methods of semi-supervised clustering are presented, including Safe semi-supervised clustering and Picture fuzzy set and picture fuzzy clustering.

2.1 Safe Semi-Supervised Clustering

Safe semi-supervised fuzzy clustering approaches, including S3FCM [11], LHC-S3FCM [12] and CS3FCM [13] are proposed by *Gan et al*. Herein, we present the fundamental knowledge of these approaches.

For S3FCM, consider the dataset $X = \{X_1, X_2, \ldots, X_k, \ldots, X_n\}$ where n is the number of data elements. C is denoted for the number of clusters. The cluster center V is defined by $\left\{V_1, V_2, \ldots, V_j, \ldots, V_C\right\}$. The membership degree of k^{th} element belonging to the i^{th} cluster is characterized by u_{ik} and m is the fuzzifier parameter. The value b_k expresses a label indicator; the value $b_k = 1$ if X_k is labeled and $b_k = 0$ otherwise. f_{ik} are the fuzzy degrees of labeled samples. The objective function of S3FCM [11] is as below:

$$J_{sa} = \sum_{k=1}^{n}\sum_{i=1}^{C} u_{ik}^{m} d_{ik}^{2} + \lambda_1 \sum_{k=1}^{n}\sum_{i=1}^{C} \left(u_{ik} - f_{ik} b_k\right)^2 d_{ik}^2 + \lambda_2 \sum_{k=1}^{n}\sum_{i=1}^{C} \left(u_{ik} - \hat{u}_{ik} b_k\right)^2 d_{ik}^2 \rightarrow Min \tag{1}$$

with: $\sum_{i=1}^{c} u_{ik} = 1, \forall k = \overline{1, n}$, $u_{ik} \in [0, 1], \forall k = \overline{1, n}$. Where λ_1 and λ_2 are the regulatory factor in which $\hat{U} = \left[\hat{u}_{ik}\right]_{c \times n}$ is the partition matrix after using FCM on unlabeled data, d_{ik} is the distance between the

k^{th} element and i^{th} cluster. The final cluster labels are determined through the algorithm [11] and the value u_{ik} is specified as follows:

$$u_{ik} = \frac{1}{1+\lambda_1+\lambda_2}\left(\frac{1+\lambda_1+\lambda_2-\sum_{j=1}^{C}\Delta_{ik}}{\sum_{j=1}^{C}\frac{d_{ik}^2}{d_{jk}^2}}+\Delta_{ik}\right) \tag{2}$$

where $\Delta_{ik} = \lambda_1 f_{ik} b_k + \lambda_2 \hat{u}_{ik} b_k$.

The below function calculates the center v_i:

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^2 x_k + \lambda_1 \sum_{k=1}^{n} (u_{ik} - f_{ik} b_k)^2 x_k + \lambda_2 \sum_{k=1}^{n} \left(u_{ik} - \hat{u}_{ik} b_k\right)^2 x_k}{\sum_{k=1}^{n} u_{ik}^2 + \lambda_1 \sum_{k=1}^{n} (u_{ik} - f_{ik} b_k)^2 + \lambda_2 \sum_{k=1}^{n} \left(u_{ik} - \hat{u}_{ik} b_k\right)^2} \tag{3}$$

On the other hand, the LHC-S3FCM [12] is expected to deal with wrong labels from additional information. The objective function is defined as follows:

$$J_{sa} = \sum_{k=1}^{n}\sum_{i=1}^{C} u_{ik}^m d_{ik}^2 + \lambda_1 \sum_{k=1}^{l}\sum_{i=1}^{C} (u_{ik} - f_{ik})^m d_{ik}^2 + \lambda_2 \sum_{k=1}^{l}\sum_{r=l+1}^{n} w_{kr} \sum_{i=1}^{C} (u_{ik} - u_{ir})^2 \rightarrow Min \tag{4}$$

with the constraints: $\sum_{i=1}^{c} u_{ik} = 1, \forall k = \overline{1,n}$

Therefore, the cluster centers v_i, the value u_{ik} for labeled data x_k and the value u_{ir} for unlabeled data x_r correspond to the below functions:

$$\mathrm{v}_i = \frac{\sum_{k=1}^{n} u_{ik}^2 x_k + \lambda_1 \sum_{k=1}^{l} (u_{ik} - f_{ik})^2 x_k}{\sum_{k=1}^{n} u_{ik}^2 + \lambda_1 \sum_{k=1}^{l} (u_{ik} - f_{ik})^2};\ \mathrm{u}_{ik} = \frac{p_{ik} + \frac{1-\sum_{j=1}^{C}\frac{p_{jk}}{q_{jk}}}{\sum_{j=1}^{C}\frac{1}{q_{jk}}}}{q_{ik}} \text{ and } \mathrm{u}_{ir} = \frac{S_{ir} + \frac{1-\sum_{j=1}^{C}\frac{s_{jr}}{t_{jr}}}{\sum_{j=1}^{C}\frac{1}{s_{jr}}}}{t_{ir}} \tag{5}$$

Another approach of FCM is Confidence-weighted Safe Semi-supervised Clustering (CS3FCM) [13] by using confidence weights. The confidence weights show various effects of samples on performance degradation. The following is the goal:

$$J_c = \sum_{k=1}^{n}\sum_{i=1}^{C} u_{ik}^m d_{ik}^2 + \lambda_1 \sum_{k=1}^{l} s_k \sum_{i=1}^{C} (u_{ik} - f_{ik})^2 d_{ik}^2 + \lambda_2 \sum_{k=1}^{l} \frac{1}{s_k} \sum_{r=l+1}^{n} w_{kr} \sum_{i=1}^{C} (u_{ik} - u_{ir})^2 \rightarrow Min \tag{6}$$

with $\sum_{i=1}^{c} u_{ik} = 1, \forall k = \overline{1,n}$; λ_1 and λ_2 are the regulatory factors. Therefore, the value of v_i, u_{ik} and u_{ir} are determined by the following functions:

$$\mathrm{v}_i = \frac{\sum_{k=1}^{n} u_{ik}^2 x_k + \lambda_1 \sum_{k=1}^{l} s_k (u_{ik} - f_{ik})^2 x_k}{\sum_{k=1}^{n} u_{ik}^2 + \lambda_1 \sum_{k=1}^{l} s_k (u_{ik} - f_{ik})^2};\ \mathrm{u}_{ik} = \frac{p_{ik} + \frac{1-\sum_{j=1}^{C}\frac{p_{jk}}{q_{jk}}}{\sum_{j=1}^{C}\frac{1}{q_{jk}}}}{q_{ik}} \text{ and } \mathrm{u}_{ir} = \frac{z_{ir} + \frac{1-\sum_{j=1}^{C}\frac{z_{ir}}{t_{ir}}}{\sum_{j=1}^{C}\frac{1}{t_{ir}}}}{t_{ir}} \tag{7}$$

The methods of Gan [11–13] (S3FCM, LHC-S3FCM, CS3FCM) achieved good clustering accuracy. However, if there may be data outliers, they would affect the determination of the final clusters.

2.2 Picture Fuzzy Set and Picture Fuzzy Clustering

By generalizing the fuzzy set in [9] and the intuitionistic fuzzy set [23], Cuong et al. introduced a definition of the picture fuzzy set [15] in 2014 and have the form as follows:

$$S = \{(x, \mu_S(x), \eta_S(x), \gamma_S(x)) \mid x \in X\} \tag{8}$$

where $\mu_S(x)$, $\eta_S(x)$ and $\gamma_S(x)$ correspond to the positive degree, the neutral degree and the negative degree of each element. And these degrees satisfy the following conditions:

$$0 \le \mu_S(x), \eta_S(x), \gamma_S(x) \le 1; 0 \le \mu_S(x) + \eta_S(x) + \gamma_S(x) \le 1 \tag{9}$$

Then, the refusal degree is computed by function:

$$\xi_S(x) = 1 - (\mu_S(x) + \eta_S(x) + \gamma_S(x)) \tag{10}$$

The objective of FC-PFS [17] aims to group the data in clusters and reduce the outliers through the concept of entropy as follows:

$$J_m(U, \eta, \xi, V) = \sum_{i=1}^{n}\sum_{j=1}^{C}\left(\mu_{ij}\left(2 - \xi_{ij}\right)\right)^m \left\|x_i - v_j\right\|^2 + \sum_{i=1}^{n}\sum_{j=1}^{C}\eta_{ij}\left(\log\eta_{ij} + \xi_{ij}\right) \to Min \tag{11}$$

with the constraints:

$$\mu_{ij}, \eta_{ij}, \xi_{ij} \in [0,1],\ \mu_{ij} + \eta_{ij} + \xi_{ij} \in [0,1], \sum_{j=1}^{C}\mu_{ij}\left(2 - \xi_{ij}\right) = 1, \sum_{j=1}^{C}\left(\eta_{ij} + \frac{1}{C}\xi_{ij}\right) = 1, i = \overline{1,n} \text{ and } j = \overline{1,C} \tag{12}$$

The values of $\mu_S(x)$, $\eta_S(x)$ and $\gamma_S(x)$ correspond to the positive, neutral and negative degrees of PFS [15]. The vector of the cluster centers denotes V.

For the above objective function, the cluster centers v_j, the membership degrees μ_{ij} and non-membership degrees η_{ij} are computed using the following formulas:

$$v_j = \frac{\sum_{i=1}^{n}\left(\mu_{ij}\left(2 - \xi_{ij}\right)\right)^m x_i}{\sum_{i=1}^{n}\left(\mu_{ij}\left(2 - \xi_{ij}\right)\right)^m} \tag{13}$$

$$\mu_{ij} = \frac{1}{\left(2 - \xi_{ij}\right)\sum_{k=1}^{C}\left(\frac{\left\|x_i - v_j\right\|^2}{\left\|x_i - v_k\right\|^2}\right)^{1/(m-1)}} \tag{14}$$

$$\eta_{ij} = \frac{\exp\left(-\xi_{ij}\right)}{\sum_{k=1}^{C}\exp\left(-\xi_{ik}\right)}\left(1 - \frac{1}{C}\sum_{k=1}^{C}\xi_{ik}\right) \tag{15}$$

In [15], the refusal degree ξ_{ij} is calculated using the Yager complement operator as follows:

$$\xi_{ij} = 1 - \left(\mu_{ij} + \eta_{ij}\right) - \left(1 - \left(\mu_{ij} + \eta_{ij}\right)^{\alpha}\right)^{1/\alpha} \tag{16}$$

where $\alpha \in (0, 1)$ is a regulatory factor and it is often chosen within [0.6–0.8].

The detailed steps for the FC-PFS algorithm are shown below.

Algorithm 1: The main steps of the FC-PFS algorithm

Input: Data set X with n elements in R^d; threshold ε; the number of clusters C; fuzzifier m; exponent $\alpha \in (0, 1)$ and the maximal number of iterations $Maxsteps > 0$.
Output: Membership matrices μ, η, ξ and cluster centers V.
1: Initialize the iteration: $t = 0$
2: $\mu_{kj}{}^{t} \leftarrow random$; $\eta_{kj}{}^{t} \leftarrow random$; $\xi_{kj}{}^{t} \leftarrow random$ $\left(k = \overline{1, n};\ \ j = \overline{1, C}\right)$ satisfy Eqs. (11)–(11)
3: **Repeat**
4: $t = t + 1$
5: Calculate V_j $(j = 1, \ldots, C)$ by Eq. (13)
6: Calculate $\mu_{kj}^{(t)}$ for labeled data $\left(k = \overline{1, n};\ \ j = \overline{1, C}\right)$ by Eq. (14)
7: Calculate $\eta_{kj}^{(t)}$ $\left(k = \overline{1, n};\ \ j = \overline{1, C}\right)$ by Eq. (15)
8: Calculate $\xi_{kj}^{(t)}$ $\left(k = \overline{1, n};\ \ j = \overline{1, C}\right)$by Eq. (16)
9: **Until** the matrices μ, η, ξ satisfy the condition $\left\|\mu^{t} - \mu^{t-1}\right\| + \left\|\eta^{t} - \eta^{t-1}\right\| + \left\|\xi^{t} - \xi^{t-1}\right\| \leq \varepsilon$ or the number of iterations reaches $Maxsteps$.

3 The Proposed Picture-Neutrosophic Trusted Safe Semi-Supervised Fuzzy Clustering

3.1 Main Ideas

The idea behind the proposed method (PNTS3FCM) is the combination between PFS and safe semi-supervised fuzzy clustering by introducing a novel objective function with four primary components. The first and the second stages are employed from the original picture fuzzy clustering method [17]. The two last stages are the semi-supervised component used to orient the clustering process by labeled and unlabeled data. The main idea is represented in Fig. 1 and the detailed steps are described in Fig. 2.

Fig. 1 illustrates the method and concept in which the input data are provided to the block of PNTS3FCM. Through the use of picture fuzzy degrees, the first step of PNTS3FCM is to reduce the distance between data components and cluster centers. The picture fuzzy set model's second step involves processing the "noisy data" by integrating the entropy quantity between the neutral and refuse degrees. The refusal degree plays an important role in reducing the effect of noise data in the objective function because of its higher value relating to noise data following [17].

To deal with "safe information", the two last stages coordinate the safe semi-supervised fuzzy clustering using both labeled and unlabeled data. PNTS3FCM has two phases: Firstly, FC-PFS is used to partition all data to get the clustering result with positive, neutral and refusal values. The second phase uses all data with these values to partition data to archive better clustering quality by enhancing safe data information and reducing noisy data.

The technique produces final clusters that are reliable and confident. We will discuss the formulation and algorithm for this concept in the next section.

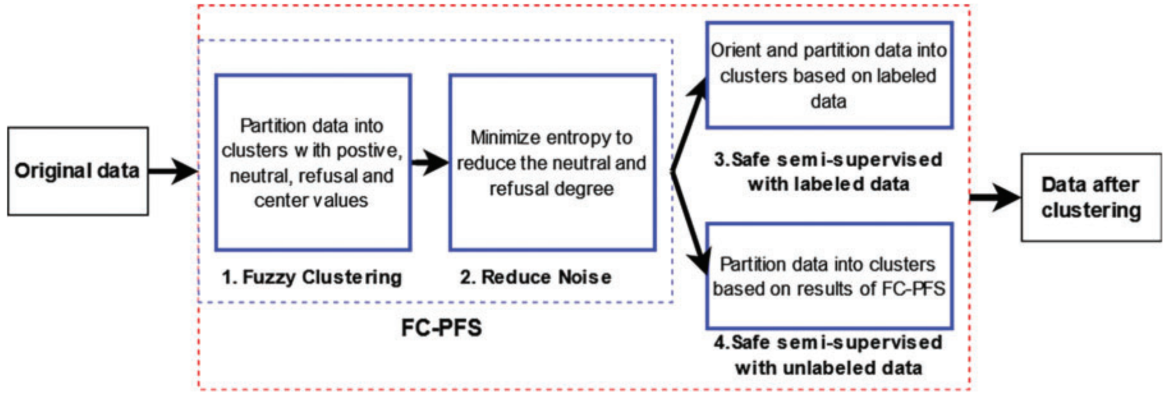

Figure 1: The main idea of the PNTS3FCM

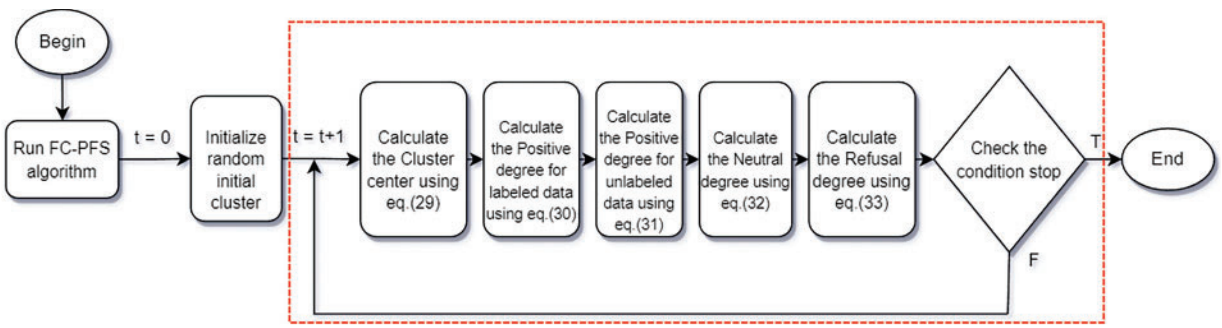

Figure 2: The details of the PNTS3FCM

3.2 Details of PNTS3FCM

As illustrated by the main idea above, this section will describe the details of the proposed model. The objective function is stated by the following formula:

$$J = \sum_{k=1}^{n}\sum_{j=1}^{C}\left(\mu_{kj}\left(2-\xi_{kj}\right)\right)^2\left\|X_k - V_j\right\|^2 + \sum_{k=1}^{n}\sum_{j=1}^{C}\eta_{kj}\left(\log\eta_{kj}+\xi_{kj}\right)$$
$$+\sum_{k=1}^{L}\sum_{j=1}^{C}\frac{\left(\mu_{kj}\left(2-\xi_{kj}\right)-f_{kj}\right)^2}{1+\left(\overline{\mu}_{kj}\left(2-\overline{\xi}_{kj}\right)-f_{kj}\right)^2}\left\|X_k - V_j\right\|^2 + \sum_{k=L+1}^{n}\sum_{j=1}^{C}\frac{\left(\mu_{kj}\left(2-\xi_{kj}\right)\right)^2}{1+\overline{\xi_{kj}}}\left\|X_k - V_j\right\|^2 \rightarrow Min \tag{17}$$

With the constraints $\left(k=\overline{1,n};\ \ j=\overline{1,C}\right)$:

$$\mu_{kj},\ \eta_{kj},\ \xi_{kj} \le 1, \sum_{j=1}^{C}\left(\eta_{kj}+\frac{\xi_{kj}}{C}\right) = 1, \text{and} \sum_{j=1}^{C}\left(\mu_{kj}\left(2-\xi_{kj}\right)\right) = 1 \tag{18}$$

where data set $X = \{X_1, X_2, \ldots, X_n\}$ having n elements, the number of labeled data in X: $L < n$; the number of clusters C; the values of positive, neutral and refusal degrees of element X_k belong to cluster j: μ_{kj}, η_{kj} and ξ_{kj}. Each part of the objective function has its own meaning. The first two parts of Eq. (17), as shown, are those of the original picture fuzzy clustering (FC-PFS) [14]. The safe semi-supervised fuzzy clustering on the picture fuzzy set is covered in the last two parts.

- **The first part** represents fuzzy clustering on the PFS.

- **The second part** represents entropy information which helps to reduce noisy data through the neutral and refusal degrees of a data point.
- **The third part** is the component for labeled data elements, where $k = 1..L$ and L is the number of labeled data elements. The numerator $\left(\mu_{kj}\left(2-\xi_{kj}\right)-f_{kj}\right)^2$ describes **semi-supervised fuzzy clustering**, in which f_{kj} is a given constant that has a value of 1 or 0.

$$f_{kj} = \begin{cases} 1 & \text{if the element k is in cluster j} \\ 0 & \text{if the element k isn't in cluster j} \end{cases} \tag{19}$$

The denominator $1+\left(\overline{\mu}_{kj}\left(2-\overline{\xi}_{kj}\right)-f_{kj}\right)^2$ describes the **safe** semi-supervised clustering. The meaning of this component is as follows: After clustering, if any data point is assigned to the correct label, the weight will be increased; otherwise, the weight will be decreased.

- Finally, **the fourth part** is the component of the unlabeled data elements, where the numerator is the same as the first part and the denominator $\left(1+\overline{\xi}_{kj}\right)$ is added to the component $\overline{\xi}_{kj}$. The meaning of this value is that after applying clustering to all data points, the denominator $\left(1+\overline{\xi}_{kj}\right)$ will be greater than 1 for unlabeled data elements with high refusal value of $\overline{\xi}_{kj}$. Indeed, the weights of these data elements are reduced.
- The **additional information** for semi-supervised fuzzy clustering is the prior picture membership degrees. We use the original FC-PFS algorithm to cluster all data, including labeled and unlabeled data. From that, we calculate four values $\left(\overline{\mu}_{kj}, \overline{\eta}_{kj}, \overline{\xi}_{kj}, \overline{V}\right)$ that guide the calculation for all data elements.

Using the Lagrangian method, the optimal solutions to the stated problem are presented in Eqs. (20)–(24) below.

$$V_j = \frac{\sum_{k=1}^{n}\left(\mu_{kj}\left(2-\xi_{kj}\right)\right)^2 X_k + \sum_{k=1}^{L}\frac{\left(\mu_{kj}\left(2-\xi_{kj}\right)-f_{kj}\right)^2}{1+\left(\overline{\mu}_{kj}\left(2-\overline{\xi}_{kj}\right)-f_{kj}\right)^2}X_k + \sum_{k=L+1}^{n}\frac{\left(\mu_{kj}\left(2-\xi_{kj}\right)\right)^2}{\left(1+\overline{\xi}_{kj}\right)^2}X_k}{\sum_{k=1}^{n}\left(\mu_{kj}\left(2-\xi_{kj}\right)\right)^2 + \sum_{k=1}^{L}\frac{\left(\mu_{kj}\left(2-\xi_{kj}\right)-f_{kj}\right)^2}{1+\left(\overline{\mu}_{kj}\left(2-\overline{\xi}_{kj}\right)-f_{kj}\right)^2} + \sum_{k=L+1}^{n}\frac{\left(\mu_{kj}\left(2-\xi_{kj}\right)\right)^2}{\left(1+\overline{\xi}_{kj}\right)^2}} \tag{20}$$

The positive degree u of the labeled data elements is

$$\mu_{kj} = \frac{f_{kj}}{\left(2-\xi_{kj}\right)\left(2+\left(\overline{\mu}_{kj}\left(2-\overline{\xi}_{kj}\right)-f_{kj}\right)^2\right)} + \\ + \frac{1-\sum_{i=1}^{C}\frac{f_{ki}}{2+\left(\overline{\mu}_{ki}\left(2-\overline{\xi}_{ki}\right)-f_{ki}\right)^2}}{\left(2-\xi_{kj}\right)\frac{\left(2+\left(\overline{\mu}_{kj}\left(2-\overline{\xi}_{kj}\right)-f_{kj}\right)^2\right)}{\left(1+\left(\overline{\mu}_{kj}\left(2-\overline{\xi}_{kj}\right)-f_{kj}\right)^2\right)}\sum_{i=1}^{C}\frac{\left\|X_k-V_j\right\|^2\left(1+\left(\overline{\mu}_{ki}\left(2-\overline{\xi}_{ki}\right)-f_{ki}\right)^2\right)}{\left\|X_k-V_i\right\|^2\left(2+\left(\overline{\mu}_{ki}\left(2-\overline{\xi}_{ki}\right)-f_{ki}\right)^2\right)}} \tag{21}$$

The positive degree u of the unlabeled data elements is

$$\mu_{kj} = \frac{1}{\left(2-\xi_{kj}\right)\sum_{i=1}^{C}\left(\frac{1+\frac{1}{\overline{\xi}_{kj}}}{1+\frac{1}{\overline{\xi}_{ki}}}\right)\frac{\left\|X_k-V_j\right\|^2}{\left\|X_k-V_i\right\|^2}} \tag{22}$$

Other degrees are shown below:

$$\eta_{kj} = \left(1-\frac{1}{C}\sum_{i=1}^{C}\xi_{ki}\right)\frac{e^{-\xi_{kj}}}{\sum_{i=1}^{C}e^{-\xi_{ki}}} \tag{23}$$

$$\xi_{kj} = \frac{1}{1 + e^{3-(\mu_{kj}+\eta_{kj})-(3-(\mu_{kj}+\eta_{kj})^{\alpha})^{\frac{1}{\alpha}}}} \tag{24}$$

Details of the FPNTS3FCM algorithm are below.

Algorithm 2: The PNTS3FCM algorithm

Input: Data set X with the number of data elements n in d dimensions, the number of labeled data in $X: L < n$; threshold ε; the number of clusters(C); fuzzifier m; exponent $\alpha \in (0, 1]$ and the maximal number of iteration $Maxsteps > 0$

Output: Membership matrices μ, η, ξ and cluster centers V.

1: Execute the FC-PFS algorithm with all data elements to get $\left(\overline{\mu}_{kj}, \overline{\eta}_{kj}, \overline{\xi}_{kj}, \overline{V}\right)$

2: Initialize the iteration: $t = 0$

3: ${\mu_{kj}}^{t} \leftarrow random$; ${\eta_{kj}}^{t} \leftarrow random$; ${\xi_{kj}}^{t} \leftarrow random$ $\left(k = \overline{1, n};\ \ j = \overline{1, C}\right)$ satisfy Eq. (18)

4: **Repeat**

5: $t = t + 1$

6: Calculate V_j $(j = 1, \ldots, C)$ by Eq. (20)

7: Calculate $\mu_{kj}^{(t)}$ for labeled data $\left(k = \overline{1, n};\ \ j = \overline{1, C}\right)$ by Eq. (21)

8: Calculate $\mu_{kj}^{(t)}$ for unlabeled data $\left(k = \overline{1, n};\ \ j = \overline{1, C}\right)$ by Eq. (22)

9: Calculate $\eta_{kj}^{(t)}$ $\left\langle\left(k = \overline{1, n};\ \ j = \overline{1, C}\right)\right\rangle$ by Eq. (23)

10: Calculate $\xi_{kj}^{(t)}$ $\left(k = \overline{1, n};\ \ j = \overline{1, C}\right)$ by Eq. (24)

11: **Until** the matrices μ, η, ξ satisfy the condition $\left\|\mu^{t} - \mu^{t-1}\right\| + \left\|\eta^{t} - \eta^{t-1}\right\| + \left\|\xi^{t} - \xi^{t-1}\right\| \leq \varepsilon$ or the number of iterations reaches $Maxsteps$.

3.3 Remarks

Advantages of the PNTS3FCM algorithm:

a) PNTS3FCM has better clustering quality than the related methods, such as FC-PFS and CS3FCM algorithm, due to the capability to handle noisy data.

b) PNTS3FCM produces more information about the clusters, such as the cluster centers and the picture fuzzy degrees (positive, neutral, negative, refusal). It deals with both "safe information" and "noisy data".

c) PNTS3FCM is the combination of three major concepts: SAFE, SEMI Clustering and PICTURE Fuzzy Set. The combination is the first trial in the literature toward practical problems.

Disadvantages of the PNTS3FCM algorithm

a) PNTS3FCM takes more computational time than the other algorithms due to the calculation of two additional parts in the objective function (24).

b) The model contains many parameters which need to be tuned in some real-world applications.

4 Experimental Results

4.1 Environmental Configuration

The experiments are performed on a Core i5-powered HP laptop using the C programming language. The selected benchmark UCI datasets [24] are described in Table 1. Outlier Detection DataSets (ODDS) [25] are given in Table 2.

Table 1: Datasets without outliers

Dataset	No. of records	No. of attributes	No. of clusters
Australian	690	14	2
Balance-scale	625	4	3
Dermatology	366	34	6
Heart	270	13	2
Iris	150	4	3
Spambase	4601	57	2
Tae	151	5	3
Waweform	5000	40	3
WDBC	569	30	2

Table 2: Datasets with outliers

Dataset	No. of samples	No. of features	No. of clusters	No. of coutlier (%)
Ecoli	336	7	8	2.6
Glass	214	9	6	4.2
Yeast	1364	8	10	4.7
Wine	178	13	3	7.7
Vertebral	310	6	3	12.5
Ionosphere	351	34	2	36

Experiments are executed to compare the proposed PNTS3FCM approach and the state-of-art methods, CS3FCM [13] and FC-PFS [17]. The classification accuracy (CA), computing time (CT) and clustering quality indicators, including DB, PBM and ASWC [26], are the criteria for evaluation. The CT is the amount of time needed to complete the computation. Value CT is computed as in (25).

$$CT = T_2 - T_1 \tag{25}$$

where T_1, T_2 is the starting time and ending time of the algorithm, respectively. The smaller value of CT reaches, the better performance of the method is. The calculation of CA [13] is given by the below equation.

$$CA = \frac{\sum_{k=1}^{n} \delta\left(y_k, map\left(\tilde{y}_k\right)\right)}{n} \tag{26}$$

where $map\left(\tilde{y}_k\right)$ is the function that determines the equivalent label for $\tilde{y}_k$ using the Kuhn–Munkres algorithm [12]. The function $\delta\left(x, y\right)$ gets two values (0 if $x \neq y$ and 1 if $x = y$). The performance of the CA index is better when it has a higher value.

The value of ASWC is computed by Eq. (27).

$$s_{xj} = \frac{b_{p,j}}{a_{p,j} + \varepsilon} \tag{27}$$

where $a_{p,j}$ is the average distance from i^{th} element to all other parts in p^{th} cluster; $b_{p,j}$ is the average distance from i^{th} element to all other elements in p^{th}cluster. ε is a tiny constant. It is added to make the denominator differ from zero (when $a_{p,j} = 0$). The higher value of the ASWC index leads to better performance.

The value of PBM [26] is determined by:

$$PBM = \left(\frac{1}{C}\frac{E_1}{E_K}D_K\right)^2 \tag{28}$$

where $E_1 = \sum_{i=1}^{n} \left\|X_i - \overline{X}\right\|$, $E_K = \sum_{j=1}^{C} \sum_{X_i \in clusterj} \left\|X_i - \overline{X}_j\right\|$, $D_K = \max_{j,l=1,\ldots,C} \left\|\overline{X}_j - \overline{X}_l\right\|$ *with* $\overline{X}_j$*is the average value of all elements in the* j^{th} *cluster*, $j = \overline{1, C}$. The higher value of the PBM index has, the better performance is.

The DB [27] is determined by (29)

$$DB = \frac{1}{C}\sum_{i=1}^{C}\left(\max_{j:\,j\neq i}\left\{\frac{S_i + S_j}{M_{ij}}\right\}\right) \tag{29}$$

where T_i is the size of i^{th} cluster. In which S_i and M_{ij} are computed by

$$S_i = \sqrt{\frac{1}{T_i}\sum_{j=1}^{T_i}\left|X_j - V_i\right|^2};\ M_{ij} = \left\|V_i - V_j\right\| \text{ with } i, j = \overline{1, C},\ \ i \neq j \tag{30}$$

The average value and standard deviation value in experimental results are denoted as Ave and STD Dev, respectively.

4.2 Experimental Results

4.2.1 Classification Accuracy

Herein, the proposed method is assessed by classification accuracy in two situations, including on all data and labeled data. Herein, the experimental results are presented following two of these cases.

Evaluation by classification accuracy on all data

Using all the data elements of 15 datasets, the classification accuracy of PNTS3FCM, FC-PFS and CS3FCM are calculated and presented as follows. Table 3 shows the classification accuracy of all data without outliers.

As shown in Table 3, PNTS3FCM gets the best results of CA on 7/9 datasets (except Australian and WDBC). FC-PFS has not achieved the highest CA on all datasets. CS3FCM is the best model on 2/9 datasets (Australian, WDBC).

From the results in Table 4, it is clear that PNTS3FCM gives correct classification results in 4 out of 6 datasets (Glass, Yeast, Vertebral, Ionosphere). The other FC-PFS is only better on the Wine dataset and CS3FCM only gives good results on the Ecoli dataset.

Table 3: Classification accuracy on all data without outliers (Bold values indicate the best results)

METHOD	PNTS3FCM		FC-PFS		CS3FCM	
	Ave	STD Dev	Ave	STD Dev	Ave	STD Dev
Australian	0.69002	0.00786	0.61856	0.00063	**0.69739**	0.00254
Balance-scale	**0.62383**	0.00556	0.51412	0.01218	0.51685	0.01939
Dermatology	**0.75452**	0.01209	0.55878	0.01647	0.64483	0.01245
Heart	**0.75298**	0.00790	0.6552	0.00195	0.7421	0.00232
Iris	**0.93085**	0.00392	0.92299	0.01076	0.89076	0.01963
Spambase	**0.84633**	0.01458	0.77682	0.00924	0.75396	0.0088
Tae	**0.48949**	0.03453	0.47875	0.00337	0.45421	0.00663
Waveform	**0.65930**	0.01140	0.55104	0.0074	0.52295	0.0076
WDBC	0.77269	0.00921	0.70639	0.0024	**0.7836**	0.0026

Table 4: Classification accuracy values on all data with outliers (Bold values indicate the best results)

METHOD	PNTS3FCM		FC-PFS		CS3FCM	
	Ave	STD Dev	Ave	STD Dev	Ave	STD Dev
Ecoli	0.51301	0.00936	0.51399	0.00509	**0.5661**	0.00974
Glass	**0.42946**	0.00741	0.42114	0.0035	0.42905	0.00625
Yeast	**0.48018**	0.02388	0.32437	0.00401	0.32909	0.0101
Wine	0.89228	0.01452	**0.95722**	0.0015	0.92186	0.00484
Vertebral	**0.51908**	0.00830	0.48733	0.00227	0.51600	0.00655
Ionosphere	**0.53823**	0.00102	0.52571	0.00037	0.53613	0.00053

Summary: During the evaluation by classification accuracy on all data, including outlier and non-outlier (15 datasets), PNTS3FCM is the best on 11 datasets (Balance-scale, Dermatology, Heart, Iris, Spambase, Tae, Waveform, Glass, Yeast, Vertebral, Ionosphere). FC-PFS is the best model on the Wine dataset. CS3FCM is the best model on three datasets (Australian, WDBC, Ecoli).

Evaluation by classification accuracy on labeled data

By using the labeled data elements of 15 datasets, the classification accuracy (CA) of PNTS3FCM, FC-PFS and CS3FCM are calculated and presented as follows. Table 5 shows the classification accuracy of labeled data without outliers.

In Table 5, PNTS3FCM gets the best results of CA on 7/9 datasets (except Iris and WDBC). FC-PFS has no highest value on all datasets. CS3FCM is the best model on 2/9 datasets (Iris, WDBC). As shown in Table 6, PNTS3FCM shows the highest values on 4/6 datasets (Ecoli, Glass, Yeast, Vertebral). FC-PFS has no highest CA on all datasets. CS3FCM is the best model on 2/6 datasets (Wine, Ionosphere).

Table 5: Classification accuracy on labeled data without outliers (Bold values indicate the best results)

METHOD	PNTS3FCM		FC-PFS		CS3FCM	
	Ave	STD Dev	Ave	STD Dev	Ave	STD Dev
Australian	**0.73497**	0.03576	0.59368	0.02729	0.73395	0.00452
Balance-scale	**0.88029**	0.07930	0.47638	0.02352	0.53585	0.01231
Dermatology	**0.69973**	0.05447	0.44638	0.04177	0.47918	0.03509
Heart	**0.76174**	0.04002	0.61693	0.03477	0.74145	0.00235
Iris	0.83296	0.02356	0.77855	0.09295	**0.85016**	0.01784
Spambase	**0.92198**	0.06753	0.69296	0.05677	0.70896	0.02728
Tae	**0.72589**	0.06706	0.53238	0.03506	0.55857	0.00782
Waveform	**0.70632**	0.05428	0.50336	0.03123	0.52359	0.0094
WDBC	0.70137	0.05464	0.6598	0.05484	**0.81087**	0.00795

Table 6: Classification accuracy on labeled data with outliers (Bold values indicate the best results)

METHOD	PNTS3FCM		FC-PFS		CS3FCM	
	Ave	STD Dev	Ave	STD Dev	Ave	STD Dev
Ecoli	**0.75089**	0.104192044	0.55551	0.05475	0.51386	0.01736
Glass	**0.64169**	0.068188815	0.43231	0.02221	0.44376	0.00876
Yeast	**0.86177**	0.10076	0.29044	0.02543	0.35129	0.01674
Wine	0.76932	0.072052299	0.80023	0.09646	**0.82272**	0.02488
Vertebral	**0.67821**	0.058126313	0.49862	0.03127	0.58464	0.01392
Ionosphere	0.52633	0.001123518	0.53435	0.01013	**0.55682**	0.00212

Summary: During the evaluation by classification accuracy on labeled data, including outlier and non-outlier (15 datasets), PNTS3FCM has better results on 11 datasets (Australian, Balance-scale, Dermatology, Heart, Spambase, Tae, Waveform, Ecoli, Glass, Yeast, Vertebral). FC-PFS has not had the highest CA on all datasets. CS3FCM is the best model on four datasets (Iris, WDBC, Wine, Ionosphere).

4.2.2 Evaluation by Clustering Quality

Summary: As in Table 7, During the evaluation clustering quality by DB index on all data, including outlier and non-outlier (15 datasets), PNTS3FCM gets the best results on ten datasets (Australian, Dermatology, Heart, Tae, Waveform, WDBC, Ecoli, Glass, Yeast, Wine). FC-PFS is the best model on three datasets (Iris, Vertebral, Ionosphere). CS3FCM is the best model on 3 datasets (Balance-scale, Spambase). This pointed out that the proposed method was better in clustering quality in not only outlier not also non-outlier data compared to others.

Table 7: The results of the DB index on all datasets (bold values indicate the best results)

METHOD	PNTS3FCM		FC-PFS		CS3FCM	
	Ave	STD Dev	Ave	STD Dev	Ave	STD Dev
Australian	**3.29466**	0.07284	3.59062	0.29431	3.80124	0.24324
Balance-scale	6.54266	0.69000	52.46293	4.3444	**5.54124**	0.22229
Dermatology	**10.50804**	0.76929	15.64693	7.57343	18.65225	3.36386
Heart	**3.12663**	0.08336	5.1217	4.26367	3.94698	0.19342
Iris	2.95875	0.02369	**2.85474**	0.0293	3.57822	0.21908
Spambase	31.26761	10.24416	33.89317	0.26046	**25.57641**	0.16618
Tae	**3.44788**	0.01591	3.8189	0.01948	3.70213	0.01689
Waweform	**14.42766**	2.92785	15.15752	3.34036	16.15447	2.87025
WDBC	**2.35134**	0.02410	2.41815	0.00587	2.83812	0.05349
Ecoli	**5.97278**	0.28630	6.49862	0.16336	8.88006	0.38829
Glass	**5.37032**	0.16483	6.62155	0.32037	6.4242	0.52104
Yeast	**10.61662**	0.58698	28.00517	2.14368	12.03848	1.11802
Wine	**2.82494**	0.01133	2.91238	0.00254	4.10052	0.10899
Vertebral	3.11417	0.02677	**3.07404**	0.0136	3.82931	0.09941
Ionosphere	2.98972	0.01699	**2.95349**	0.01212	3.3906	0.06444

4.2.3 Evaluation by Computational Time (in seconds)

We compare PNTS3FCM and CS3FCM on 15 datasets using computational time. Table 8 shows the results of evaluation clustering quality by computational time on data without outlier datasets.

Table 8: The computational time on all datasets (Bold values indicate the best results)

METHOD	PNTS3FCM		CS3FCM	
	Ave	STD Dev	Ave	STD Dev
Australian	**0.16610**	0.01027	0.32286	0.02052
Balance-scale	**0.11395**	0.00931	1.20235	0.00396
Dermatology	**0.91512**	0.11784	1.38068	0.05733
Heart	0.06533	0.00100	**0.06518**	0.00315
Iris	**0.02343**	0.02343	0.02542	0.00312
Spambase	**2.53490**	0.18252	3.16697	0.87895
Tae	0.04673	0.00202	**0.03258**	0.00480
Waweform	**3.54067**	0.09443	8.08507	4.26694
WDBC	**0.20360**	0.00334	0.39131	0.00786
Ecoli	1.71820	0.14015	**0.92510**	0.06711
Glass	0.57863	0.02675	**0.41492**	0.01662
Yeast	11.35089	0.41673	**6.10282**	0.80203

(Continued)

Table 8: Continued

METHOD	PNTS3FCM		CS3FCM	
	Ave	STD Dev	Ave	STD Dev
Wine	0.05700	0.00268	**0.05651**	0.00182
Vertebral	**0.11455**	0.00255	0.18393	0.00592
Ionosphere	**0.20543**	0.00606	0.45496	0.01373

Summary: During the evaluation of computational time on all data, including outlier and non-outlier (15 datasets), PNTS3FCM has better results on nine datasets. CS3FCM is the best model on six datasets. The proposed method seems to be better with a more significant number of data clusters. To get better results, the proposed method is firstly based on Picture fuzzy set that has more information to reduce the noise or hesitation in partitioning data. Secondly, PNTS3PFCM has a safe semi-supervised part for labeled and unlabeled data that can cope with the doubt labeled data, then reduce their effectiveness in the clustering process.

5 Conclusion

This research suggested a novel technique called Picture-Neutrosophic Trusted Safe Semi-Supervised Fuzzy Clustering (PNTS3FCM) to address the issue of data clustering with high confidence and noisy information. PNTS3FCM is constructed based on combining Picture Fuzzy Sets, Neutrosophic Sets and safe fuzzy semi-supervised clustering (PFS). This method consists of 4 critical parts: the clustering portion, the outlier solution part and the safe semi-supervised fuzzy clustering with labeled and unlabeled data. Through the use of PFS, the first stage of PNTS3FCM aims to reduce the distance between data components and cluster centers. The model's second step involves processing the "noisy data" by integrating the entropy quantity between the neutral and refuse degrees. The third and fourth stages coordinate the safe semi-supervised fuzzy clustering using both labeled and unlabeled data to solve the safety information. We also provide an iterative technique from the formulation to construct the cluster centers and memberships. The method produces final clusters that are reliable and confident.

PNTS3FCM has illustrated its effectiveness by comparing it with two related methods, including FC-PFS and CS3FCM algorithm. The experiment results show that PNTS3FCM is better than the others in terms of computational time and clustering quality. Even though the proposed PNTS3FCM mainly focuses on eliminating or reducing noisy data elements, this method still has some limitations. First of all, PNTS3FCM takes a long time to compute. Secondly, it needs an increased number of parameters. In the future, an effective optimization algorithm will be studied and introduced to overcome these limitations.

Acknowledgement: We are grateful for the support from the staff of the Institute of Information Technology, Vietnam Academy of Science and Technology.

Funding Statement: This research is funded by Graduate University of Science and Technology under grant number GUST.STS.ĐT2020-TT01.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] X. Ji, S. Liu, P. Zhao, X. Li and Q. Liu, "Clustering ensemble based on sample's certainty," *Cognitive Computation*, vol. 13, no. 4, pp. 1034–1046, 2021.

[2] J. Zhang, H. Wang, S. Huang, T. Li, P. Jin *et al.,* "Co-adjustment learning for co-clustering," *Cognitive Computation*, vol. 13, no. 2, pp. 504–517, 2021.

[3] P. H. Thong, "Some novel hybrid forecast methods based on picture fuzzy clustering for weather nowcasting from satellite image sequences," *Applied Intelligence*, vol. 46, no. 1, pp. 1–15, 2017.

[4] A. Khosravanian, M. Rahmanimanesh, P. Keshavarzi and S. Mozaffari, "Fuzzy local intensity clustering (FLIC) model for automatic medical image segmentation," *The Visual Computer*, vol. 37, no. 5, pp. 1185–1206, 2021.

[5] S. A. Kumar, B. S. Harish and V. M. Aradhya, "A picture fuzzy clustering approach for brain tumor segmentation," in *2016 Second Int. Conf. on Cognitive Computing and Information Processing (CCIP)*, Mysuru, India, pp. 1–6, 2016.

[6] G. Bode, T. Schreiber, M. Baranski and D. Müller, "A time series clustering approach for building automation and control systems," *Applied Energy*, vol. 238, no. 11, pp. 1337–1345, 2019.

[7] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*. New York: Plenum Press. http://dx.doi.org/10.1007/978-1-4757-0450-1.

[8] N. Grira, M. Crucianu and N. Boujemaa, "Active semi-supervised fuzzy clustering," *Pattern Recognition*, vol. 41, no. 5, pp. 1834–1844, 2008.

[9] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.

[10] P. H. Thong and L. H. Son, "An overview of semi-supervised fuzzy clustering algorithms," *International Journal of Engineering and Technology*, vol. 8, no. 4, pp. 301–306, 2016.

[11] H. Gan, "Safe semi-supervised fuzzy C-means clustering," *IEEE Access*, vol. 7, pp. 95659–95664, 2019.

[12] H. Gan, Y. Fan, Z. Luo and Q. Zhang, "Local homogeneous consistent safe semi-supervised clustering," *Expert Systems with Applications*, vol. 97, pp. 384–393, 2018.

[13] H. Gan, Y. Fan, Z. Luo, R. Huang and Z. Yang, "Confidence-weighted safe semi-supervised clustering," *Engineering Applications of Artificial Intelligence*, vol. 81, pp. 107–116, 2019.

[14] L. Guo, H. Gan, S. Xia, X. Xu and T. Zhou, "Joint exploring of risky labeled and unlabeled samples for safe semi-supervised clustering," *Expert Systems with Applications*, vol. 176, pp. 114796–114803, 2021.

[15] B. C. Cuong and V. Kreinovich, "Picture fuzzy sets," *Journal of Computer Science and Cybernetics*, vol. 30, no. 4, pp. 409–420, 2014.

[16] F. Smarandache, *Neutrosophy: Neutrosophic probability, set and logic: analytic synthesis & synthetic analysis*. Santa Fe, Rehoboth, MA, USA: American Research Press, 1998.

[17] P. H. Thong and L. H. Son, "Picture fuzzy clustering: A new computational intelligence method," *Soft Computing*, vol. 20, no. 9, pp. 3549–3562, 2016.

[18] J. Zhou, C. P. Chen, L. Chen and H. X. Li, "A collaborative fuzzy clustering algorithm in distributed network environments," *IEEE Transactions on Fuzzy Systems*, vol. 22, no. 6, pp. 1443–1456, 2013.

[19] L. H. Son, "DPFCM: A novel distributed picture fuzzy clustering method on picture fuzzy sets," *Expert Systems with Applications*, vol. 42, no. 1, pp. 51–66, 2015.

[20] C. Wu and Y. Chen, "Adaptive entropy weighted picture fuzzy clustering algorithm with spatial information for image segmentation," *Applied Soft Computing*, vol. 86, no. 4, pp. 105888–105927, 2020.

[21] C. Wu and Z. Kang, "Robust entropy-based symmetric regularized picture fuzzy clustering for image segmentation," *Digital Signal Processing*, vol. 110, no. 1, pp. 102905–102933, 2021.

[22] C. Wu and N. Liu, "Suppressed robust picture fuzzy clustering for image segmentation," *Soft Computing*, vol. 25, no. 5, pp. 3751–3774, 2021.

[23] K. Atanassov, "Intuitionistic fuzzy sets," *International Journal Bioautomation*, vol. 20, pp. S1–S6, 2016.

[24] UCI Machine learning repository, "Data," 2021. [Online]. Available: https://archive.ics.uci.edu/ml/index.php.

[25] Outlier detection datasets (ODDS), "Data," 2016. [Online]. Available: http://odds.cs.stonybrook.edu/.

[26] L. Vendramin, R. J. Campello and E. R. Hruschka, "Relative clustering validity criteria: A comparative overview," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 3, no. 4, pp. 209–235, 2010.