Journal of Computer Science and Cybernetics, V.38, N.1 (2022), 47–61
DOI no. 10.15625/1813-9663/38/1/16720

AN IMPROVEMENT OF TRUSTED SAFE SEMI-SUPERVISED FUZZY CLUSTERING METHOD WITH MULTIPLE FUZZIFIERS

TRAN MANH TUAN[1], PHUNG THE HUAN[2], PHAM HUY THONG[3], TRAN THI NGAN[1,*], LE HOANG SON[3]

[1] *Thuyloi University, Hanoi, Vietnam*

[2] *University of Information and Communication Technology, Thai Nguyen University, Vietnam*

[3] *VNU Information Technology Institute, Vietnam National University, Hanoi, Vietnam*

Abstract. Data clustering is applied in various fields such as document classification, dental X-ray image segmentation, medical image segmentation, etc. Especially, clustering algorithms are used in satellite image processing in many important application areas, including classification of vehicles participating in traffic, logistics, classification of satellite images to forecast droughts, floods, forest fire, etc. In the process of collecting satellite image data, there are a number of factors such as clouds, weather, etc. that can affect to image quality. Images with low quality will make the performance of clustering algorithms decrease. Apart from that, the parameter of fuzzification in clustering algorithms also affects to clustering results. In the past, clustering methods often used the same fuzzification parameter, $m = 2$. But in practice, each element should have its own parameter m. Therefore, determining the parameters m is necessary to increase fuzzy clustering performance. In this research, an improvement algorithm for the data partition with confidence problem and multi fuzzifier named as TS3MFCM is introduced. The proposed method consists of three steps namely as "FCM for labeled data", "Data transformation", and "Semi-supervised fuzzy clustering with multiple point fuzzifiers". The proposed TS3MFCM method is implemented and experimentally compared against with the Confidence-weighted Safe Semi-Supervised Clustering (CS3FCM). The performance of proposed method is better than selected methods in both computational time and clustering accuracy on the same datasets.

Keywords. Fuzzy clustering; Semi-supervised fuzzy clustering; Safe semi-supervised fuzzy clustering; Multiple fuzzifiers.

1. INTRODUCTION

Data clustering divides objects into different groups with the high similarity of elements in each group [1, 2]. Data clustering algorithms are separated into two subgroups, including

*Corresponding author.
E-mail addresses: tmtuan@tlu.edu.vn (T.M.Tuan), thongph@vnu.edu.vn (P.H.Thong), pthuan@ictu.edu.vn (P.T.Huan), ngantt@tlu.edu.vn (T.T.Ngan), sonlh@vnu.edu.vn (L.H.Son)

hard and fuzzy clustering. In the former one, each data point belongs to a unique cluster. In the later one, each data point can belong to many different clusters with a specific probability. The popular fuzzy clustering algorithm proposed by Bezdek [3] is Fuzzy C-Means (FCM) method. FCM performed by optimizing the distances among the data points and centers of corresponding clusters. FCM takes advantages of the flexibility in fuzzy logic [4]. Fuzzy clustering algorithms and its extensions are applied in many different applications [5–8]. To get the higher accuracy, some additional information was added to the clustering process. Then, the fuzzy clustering algorithms are called the semi-supervised fuzzy clustering algorithms. Semi-Supervised Fuzzy C-Means (SSFCM) method [9] is one of the most popular algorithms. The objective function of SSFCM consists of components, corresponding to the integration between unsupervised learning and supervised learning. Thus, many improvements of SSFCM were introduced to deal with various problems [10–12]. In semi-supervised fuzzy clustering algorithm, when a part of data is labeled, some of labeled data could be clustered incorrectly. To deal with this issue, safe semi-supervised fuzzy clustering method (CS3FCM) is proposed by Gan [13]. CS3FCM is based on the confidence-weight of each sample to get high clustering performance. Semi-supervised clustering algorithms aim to get two targets [14–17]. The first one is to cluster data and to label each data then. The second one is to improve clustering quality using available knowledge. To improve clustering performance, most of the methods change the formula of the objective function. The value of fuzzy parameter is often chosen as a constant. Typically, in FCM, SSFCM, and CS3FCM algorithms, the value of parameter $m = 2$ is fixed. Fuzzy parameter represents the uncertainty of each data element. Thus, the consideration on determining the different values of m for each data element is necessary to increase the performance of fuzzy clustering algorithms. Fuzzy parameter represents the uncertainty of each data element. Thus, the consideration on determining the different values of m for each data element is necessary to increase the performance of fuzzy clustering algorithms. There were many studies on extending the fuzzy parameter by defining a fuzzy value $m \in [m_1, m_2]$ for each iteration [18]. Khang, T. D. et al. [19] proposed an improvement of FCM with different fuzzifiers for each element in the dataset. The fuzzifier of a specific element was calculated based on the distribution among that element and surrounding ones. The main idea of this method is using multiple fuzzifiers instead of unique fuzzifier in the FCM. With the dataset $X = \{X_1, ..., X_N\}$ number of clusters C, membership degree u_{ij} of element i^{th} in j^{th} cluster, the distance d_{ij} from i^{th} data element to cluster center V_j, the fuzzifiers (m_i) of i^{th} data element can be defined as follows

$$m_i = m_1 + (m_2 - m_1)\left(\frac{S_i - S_{\min}}{S_{\max} - S_{\min}}\right)^{\alpha};\ i = \overline{1, N}, \tag{1}$$

where m_1, m_2 is the lower and upper boundary of m_i ($1 \leq m_1 \leq m_2$), α is an exponent parameter, $S_i = \sum_{j=1}^{N/C} \delta_{ij}$; $\delta_{ij} = \|X_i - X_j\|$, $\left(\forall i,\ j = \overline{1, N}\right)$; $S_{\max} = \max_{i=1,...,N}(S_i)$; $S_{\min} = \min_{i=1,...,N}(S_i)$.

Using the values of m_i, the diagram of Fuzzy C-Mean Clustering with Multiple Fuzzifiers (MC-FCM) algorithm is shown in [19].

Another factor that affects to the performance of clustering process is noise and outliers in data. In many problems, data may contain incorrect information or noises. For example, in collecting of ship satellite images [17], due to the shooting angle or confounding factors such as clouds, fog, etc., the obtained images may contain noises. Thus, when applying processing

techniques, ships can be misidentified as islands or lighthouses, etc. Process of dealing with incorrect data and noisy data is called the data partition with confidence problem, including "safe information" and "noisy data". The objective of the data clustering problem with confidence can be stated that by using data clustering, the unlabeled data points will be properly labeled of clusters and incorrect labeled data points will be relabeled exactly. It means that to find the "best" boundary between correctly labeled data points and incorrectly labeled data points.

The main idea to solve the data partition with confidence problem as depicted in the researches of Gan et al. [13, 20-21] lies in two principal steps: i) to compute the confidence weights of labeled data with a local graph W; ii) to formulate and determine the cluster centers V and fuzzy membership values N_c according to the labeled data having high confidence weights. The data used in Confidence-weighted safe semi-supervised clustering (CS3FCM) [13] include both of labeled data $X = [\mathrm{x}_1,...,\mathrm{x}_l]$ and unlabeled data $X_u = [\mathrm{x}_{l+1},...,\mathrm{x}_n]$. The different data elements have different effects to clustering performance. In this method, Gan et al. used FCM to divide all data elements into C clusters then calculated the partition matrix $\widetilde{U} = [\widetilde{u}]_{c\times n}$ and estimated the output labels $\widetilde{Y} = [\widetilde{y}_1,...,\widetilde{y}_l,\widetilde{y}_{l+1},...,\widetilde{y}_n]$ using Kuhn-Munkres algorithm [22]. For each labeled element x_k, the safe confidence s_k is defined.

The dataset with noise data, outliers and misclassified data is illustrated as in Figure **1** below.

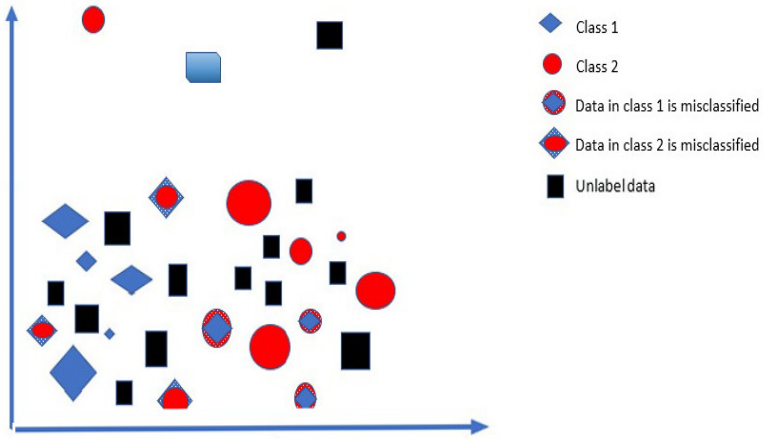


Figure 1: The dataset with a part of data labeled

In Figure 1, a part of data set is labeled. However, the uncertainty of each data element is different. This is represented by the size of data points in the figure. Moreover, some of labeled data points are misclassified. This happens when a data point in Class 1 is assigned as Class 2 or a data point in Class 2 is assigned as Class 1. Apart from that, there are some outliers existing in the dataset. Then, it is necessary to check the accuracy of labeled data. The effects of each data point to classification or clustering accuracy are different and have to be defined. This can be evaluate by using multiple fuzzifiers.

In this paper, an improvement algorithm for the data partition with confidence problem using multiple fuzzifiers named as TS3MFCM is introduced. This method reconciles labeled

data using modified FCM with the weights of labeled and unlabeled data neighbors instead of working on the whole dataset as in [13]. **The differences** of TS3MFCM comparing with CS3FCM and MC-FCM are given as below:

i. After apply modified FCM, **the labeled data with small impact is either set up with very low membership values or removed from the set of labeled data** while CS3FCM uses all labeled data in clustering process.

ii. The cluster centers obtained by applying modified FCM are used in order to compute the membership values of the unlabeled data. Thus, membership values of labeled and unlabeled data are contained in prior membership degrees ($\bar{U}$). Thus, **the additional information in TS3MFCM is a mixture of labeled data and the prior membership degrees ($\bar{U}$)** while CS3FCM only uses labeled data as additional information.

iii. **TS3MFCM uses multiple fuzzifiers for each data point to control the data clustering process**. In this step, the prior membership degrees ($\bar{U}$) are used to assist clustering progress in generating the final cluster centers and membership values for all data points. **We use a semi-supervised fuzzy clustering with multiple fuzzifiers method in order to partition the whole dataset** with the initial membership ($\bar{U}$) instead of the normal fuzzy clustering method in MC-FCM.

The proposed TS3MFCM method is implemented on specific datasets and experimentally compared with the CS3FCM. By these experiments, TS3MFCM is better than selected methods on the same datasets in computational time and clustering accuracy results.

The rests of this paper are structured as follows. The TS3MFCM method is presented in Section 2. The experimental results of implementing TS3MFCM and CS3FCM on six different datasets are given in Section 3. We draw conclusions and highlight further studies in the last section.

2. THE PROPOSED METHOD

2.1. Main idea of TS3MFCM

TS3MFCM consists of 3 following steps:

Step 1. (FCM for labeled data)
Using an improved algorithm of FCM to divide the original data points into clusters, with new weights based on labeled and unlabeled neighbors.

Step 2. (Data transformation process)
The cluster centers, obtained in Step 1, are used to compute the membership degrees of unlabeled data points. The values of membership in both labeled and unlabeled data will produce the prior membership degrees ($\bar{U}$) for the next step.

Step 3. (Semi-supervised fuzzy clustering with multiple point fuzzifiers)
Using a semi-supervised fuzzy clustering algorithm with multiple fuzzifiers to control the data clustering process.

The framework of TS3MFCM algorithm is given in Figure **2** as follows.

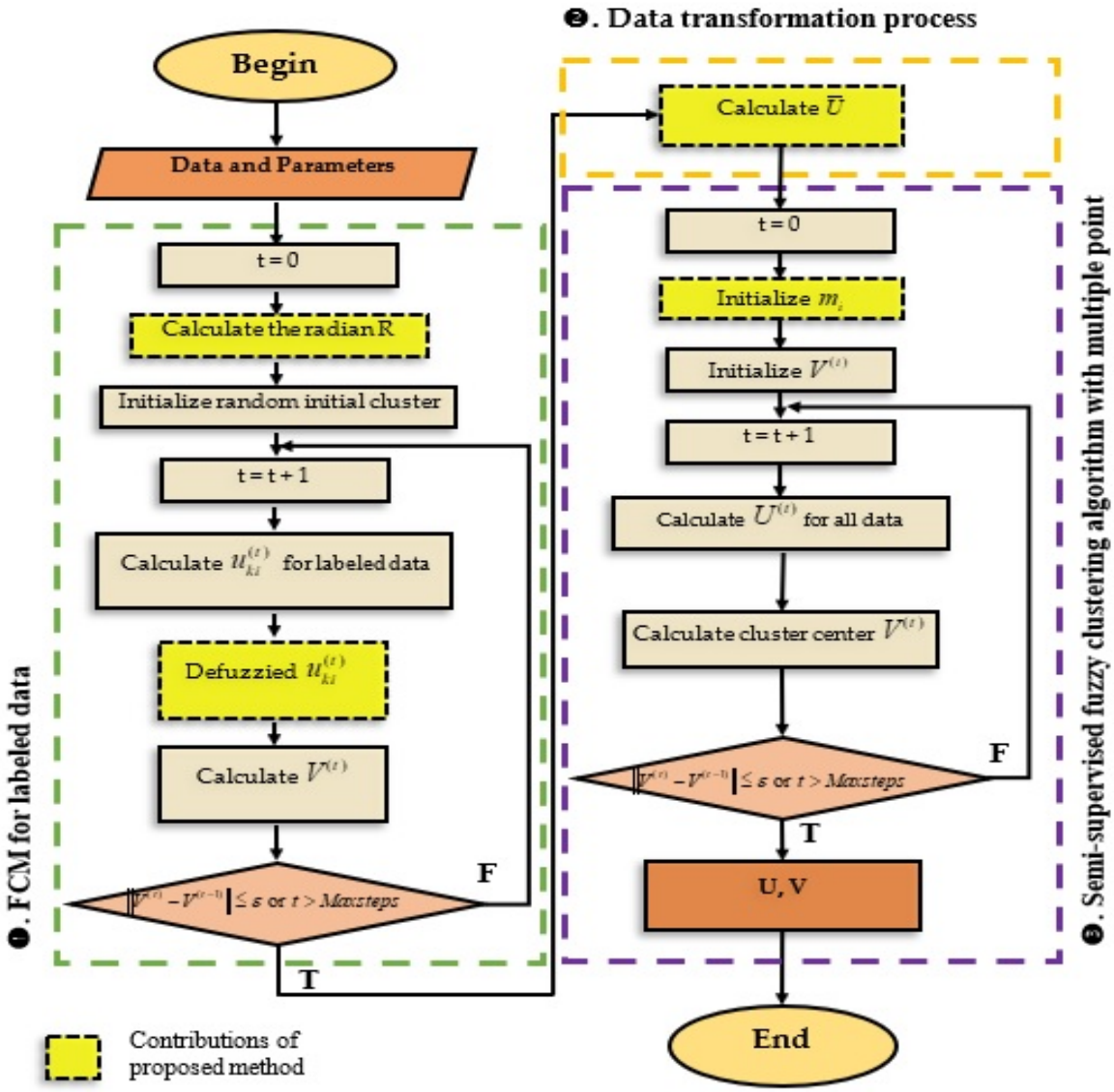


Figure 2: The flowchart of TS3MFCM algorithm

2.2. Details of the TS3MFCM

2.2.1. Step 1 (FCM for labeled data)

In this step, the algorithm compares the labeled data elements to determine the data elements with high and low confidence. To do this, we modify the original FCM algorithm with the new objective as follows

$$J = \sum_{k=1}^{L} \sum_{i=1}^{C} \frac{n_{1k} + n_{2k}}{n_{3k} + 1} u_{ki}^{m} d_{ki}^{2} \to Min, \tag{2}$$

$$u_{ki} \in [0,1]\,; k = 1, ..., L,\ i = 1, ..., C, \tag{3}$$

$$\sum_{i=1}^{C} u_{ki} = 1; k = 1, ..., L, \tag{4}$$

where n_{1k}, n_{2k}, n_{3k} are the number of unlabeled data neighbors, the number of neighbors with the same label to x_k, the number of neighbors with different label to x_k, respectively. These neighbors are defined by using Euclidean distance based on the radius R. The value of R is calculated as $(d_{\max} - d_{\min})/10$ where $d_{\min}$, $d_{\max}$ are the smallest and largest distances between two ubiquitous data points. The symbols L, C and d_{ki} are denoted for the number of labeled data, the number of clusters, and the distance between k^{th} data point and i^{th} cluster center. The cluster centers and membership degrees are computed as below.

$$V_i = \frac{\sum\limits_{k=1}^{L} \frac{n_{1k}+n_{2k}}{n_{3k}+1} u_{ki}^m X_k}{\sum\limits_{k=1}^{L} \frac{n_{1k}+n_{2k}}{n_{3k}+1} u_{ki}^m};\ i = 1, 2, ..., C, \tag{5}$$

$$u_{ki} = \frac{1}{\sum\limits_{j=1}^{C} \left(\frac{d_{ki}}{d_{kj}}\right)^{\frac{2}{m-1}}};\ k = 1, .., L,\ i = 1, 2, ..., C. \tag{6}$$

For each incorrectly labeled data point, we use defuzzification technique to reduce its membership value. If assigned cluster is different from the label of that data point, the membership value u_{ki} is reduced using (7).

$$u_{ki} = \begin{cases} \frac{u_{ki}}{2} & \text{if label of cluster } i \text{ is same to label of } x_k, \\ u_{ki} + \frac{u_{kj}}{2(C-1)} & \text{if } i \neq j \text{ and label of cluster } j \text{ is same to label of } x_k. \end{cases} \tag{7}$$

So that, the labeled data point with small impact is either set a very low membership value or removed from the set of labeled data. The modified FCM algorithm is shown in Algorithm 1 below.

2.2.2. Step 2 (Data transformation)

This is the transfer step between Step 1 and Step 3 (below). From the outputs of Step 1, we gather the cluster centers V of labeled data and use them as the initial cluster centers for the dataset of unlabeled data points. Membership values of both labeled and unlabeled data will produce the prior membership degrees ($\bar{\mathrm{U}}$) for the method in next step. Thus, in our approach, the pre-defined information of the semi-supervised fuzzy clustering is a mixture of the prior membership degrees ($\bar{\mathrm{U}}$) and labeled data.

2.2.3. Step 3 (Semi-supervised fuzzy clustering algorithm with multiple point fuzzifiers)

A semi-supervised fuzzy clustering with the fuzzifier for each point of data and the prior membership values ($\bar{\mathrm{U}}$) for all data points is introduced. This algorithm is named

Algorithm 1. The main steps of the modified FCM algorithm

Input: Data set X; the number of labeled data points in X : $L < N$; exponent α; C; ε; m and $Maxsteps$.

Output: Membership matrices u and cluster centers V.

BEGIN

1: Set $t = 0$

2: Initialize original cluster centers: $V_i^{(t)} \leftarrow random;\ \ i = 1, \ldots, C$

//Repeat **3-7**:

3: $t = t + 1$

4: Calculate $u_{ki}^{(t)}$ for labeled data ($k = 1, ..., L;\ \ i = 1, ..., C$) by (6).

5: Defuzzied $u_{ki}^{(t)}$ according to (7).

6: Calculate $V_i^{(t)}$ ($i = 1, ..., C$) using (5).

7: Check the stop condition: $\left\|V^{(t)} - V^{(t-1)}\right\| < \varepsilon$ or $t > MaxStep$. If this condition is satisfied, the algorithm is stop. Otherwise, return **3**.

END

as semi-supervised fuzzy clustering algorithm with multiple point fuzzifiers (MCSSFC-P). MCSSFC-P has objective function defined by

$$\mathrm{J(U,V)} \ = \sum_{\mathrm{i=1}}^{\mathrm{N}} \sum_{\mathrm{j=1}}^{\mathrm{C}} |u_{\mathrm{ij}} - \overline{u}_{\mathrm{ij}}|^{\mathrm{m_i}} * \|X_i - V_j\|^2 \rightarrow Min \tag{8}$$

with the constraints

$$\begin{array}{l} u_{ij} \in [0,1],\ \ \sum_{j=1}^{C} u_{ij} = 1, \quad \forall i = \overline{1,N}, \\ \overline{u}_{ij} \in [0,1],\ \ \sum_{j=1}^{C} \overline{u}_{ij} \leq 1, \quad \forall i = \overline{1,N}, \end{array} \tag{9}$$

where the dataset $X = \{X_1, ..., X_N\}$; The number of clusters C, the membership degree (u_{ij}) of i^{th} element in cluster j^{th}, the distance d_{ij} from data element i^{th} to cluster center V_j. The prior membership degree matrix ($\bar{\mathrm{U}}$) is defined via the output of Step 2. In order to solve optimal problem (8-9), Lagrange multiplier method with Lagrange function in (10) is applied

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{j=1}^{C} |u_{ij} - \bar{u}_{ij}|^{m_i} \|X_i - V_j\|^2 - \sum_{i=1}^{N} \lambda_i \left(\sum_{j=1}^{C} u_{ij} - 1 \right). \tag{10}$$

For the constant values of m_i, taking the derivative of $\mathcal{L}$ by V_j we get

$$\frac{\partial \mathcal{L}}{\partial V_j} = -2 \sum_{i=1}^{N} |u_{ij} - \bar{u}_{ij}|^{m_i} * \|X_i - V_j\|, \ \ j = \overline{1,C}. \tag{11}$$

Let the derivative in (11) be zero, cluster centers V_j are defined by

$$V_j = \frac{\sum_{i=1}^{N} |u_{\mathrm{ij}} - \overline{u}_{ij}|^{m_i} X_i}{\sum_{i=1}^{N} |u_{\mathrm{ij}} - \overline{u}_{ij}|^{m_i}}, \ \ j = \overline{1,C}. \tag{12}$$

For the constant values of m_i, taking the derivative of $\mathcal{L}$ by u_{ij}, we get

$$\frac{\partial \mathcal{L}}{\partial u_{ij}} = m_i(u_{ij} - \bar{u}_{ij})^{m_i - 1} \|X_i - V_j\|^2 - \lambda_i, \;\; i = \overline{1, N}, \; j = \overline{1, C}. \tag{13}$$

Let the derivation in (13) be zero, membership degrees u_{ij} are calculated using (14)

$$u_{\text{ij}} = \overline{u}_{ij} + \left(1 - \sum\nolimits_{k=1}^{C} \overline{u}_{ik}\right) \frac{\left(\frac{1}{d_{\text{ij}}}\right)^{\frac{2}{m_i - 1}}}{\sum_{k=1}^{C} \left(\frac{1}{d_{ik}}\right)^{\frac{2}{m_i - 1}}}, \;\; i = \overline{1, N}, \; j = \overline{1, C}, \tag{14}$$

where $d_{\text{ij}} = \|X_i - V_j\|$, $d_{\text{ik}} = \|X_i - V_k\|$.

In MCSSFC-P algorithm, input data and parameters are used to define fuzzifiers for each data sample by equation (1). Using these fuzzifiers, a semi-supervised fuzzy clustering algorithm is implemented to determine the cluster centers and membership degree matrix. The detail of MCSSFC-P is given in Algorithm **2** below.

Algorithm 2. The main steps of MCSSFC-P

Input: Data set X; C; $\overline{U}$; m_i,$(i = \overline{1, N})$; ε; $Maxsteps$.

Output: U and V.

BEGIN

1: Set $t = 0$

2: Initialize m_i using (1).

3: Initialize original cluster centers: $V_i^{(t)} \leftarrow random$; $i = 1, \ldots, C$

//Repeat **4-7**:

4: $t = t + 1$

5: Compute U^t using (14).

6: Compute V^t using (12).

7: Check the stop condition: $\left\|V^{(t)} - V^{(t-1)}\right\| < \varepsilon$ or $t > MaxStep$. If this condition is satisfied, the algorithm is stop. Otherwise, return **4**.

END

3. EXPERIMENTAL RESULTS

3.1. Environmental setup

The algorithms, including MC-FCM, CS3FCM and TS3MFCM, are implemented on HP laptop with Core i5 processor, using DevC++ programming language. The datasets are taken from the Outlier Detection DataSets [23] demonstrated in Table 1 and Airbus Ship Detection Challenge dataset [24] demonstrated in Table 2.

From the Airbus Ship Detection Challenge dataset, we use 20 satellite ship images demonstrated in Table **2** below.

Table 1: Datasets with outlier data

Dataset	**No. of samples**	**No. of features**	**No. of class**	**No. of outlier %**
Dermatology	366	34	6	2.1
Ecoli	1364	8	10	4.7
Glass	214	9	6	4.2
Ionosphere	351	34	2	36.0
Vertebral	310	6	3	12.5
Wdbc	569	30	2	5.6

Table 2: Data description of ship satellite images

Class	**No. of samples**	**No. of ships**	**No. of island**	**Size (pixel)**
Class 1	5	5	2	768× 768
Class 2	5	6	1	768× 768
Class 3	5	7	2	768× 768
Class 4	5	8	2	768× 768

Criteria for evaluation are classification accuracy (CA), clustering quality by DB index and computational time (CT). The classification accuracy (CA) [13] for the semi-supervised clustering methods is as,

$$CA = \frac{\sum_{k=1}^{n} \delta(y_k, \mathrm{map}(\widetilde{y}_k))}{n}, \tag{15}$$

where the function $\delta(x, y)$ has value of 1 if $x = y$ and 0 if $x \neq y$, $\mathrm{map}(\widetilde{y}_k)$ is the function that maps $\widetilde{y}_k$ to an equivalent label using the Kuhn–Munkres algorithm [22]. The maximum value indicates the better performance for CA index. The unit of calculation is percentage (%). In addition, an internal clustering quality index DB [26] was used to assess the ratio involving within-group and between-group distances. The DB clustering quality [26] is shown in equation (16)

$$DB = \frac{1}{C} \sum_{i=1}^{C} \left(\max_{j:j\neq i} \left\{ \frac{S_i + S_j}{M_{ij}} \right\} \right), \tag{16}$$

$$S_i = \sqrt{\frac{1}{T_i} \sum_{j=1}^{T_i} \|X_j - V_i\|^2}, \tag{17}$$

$$M_{ij} = \|V_i - V_j\|, \ (i, j = \overline{1, C}, \ \ i \neq j). \tag{18}$$

The minimum value indicates the better performance for DB index.

The CT is the length of time required to perform a computational process in equation (19)

$$CT = T_2 - T_1, \tag{19}$$

where T_2 is the ending time, and T_1 is the starting time to run the algorithm. The minimum value indicates the better performance for CT index. The unit of calculation is seconds (s).

The proposed TS3MFCM method is experimentally compared with CS3FCM algorithm [13]. The validity indices in these implementation are classification accuracy, clustering quality and computational time.

3.2. Results

3.2.1. Evaluation on outlier detection datasets

Using all the data elements in selected datasets, the classification accurac, clustering quality and computational time of TS3MFCM and CS3FCM are calculated and showed in Table 3.

Table 3: The values of validity indices on all data with outliers (Bold values indicate the best ones in given dataset)

CRITERIA	**Classification accuracy**		**Clustering quality**		**Computational time (s)**	
METHOD	TS3MFCM	CS3FCM	TS3MFCM	CS3FCM	TS3MFCM	CS3FCM
Dermatology	**0.92** ±0.01	0.64 ±0.01	**3.44** ±1.96	18.65 ±3.36	1.38 ±0.05	**1.27** ±0.06
Ecoli	**0.44** ±0.01	0.42 ±0.00	**6.42** ±0.52	19.81 ±3.56	0.89 ±0.27	**0.42** ±0.01
Glass	0.54 ±0.01	**0.57** ±0.01	**8.88** ±0.39	29.23 ±6.93	**0.30** ±0.10	0.92 ± 0.02
Ionosphere	**0.62** ± 0.01	0.54 ±0.00	**3.39** ±0.06	8.95 ±1.08	1.52 ±0.37	**0.46** ±0.01
Vertebral	**0.66** ±0.01	0.52 ±0.01	**3.83** ±1.10	5.89 ±1.24	**0.09** ± 0.00	0.18 ±0.01
Wdbc	**0.65** ±0.00	0.59 ±0.01	**2.84** ±0.05	3.92 ±0.11	0.73 ±0.10	**0.39** ±0.02

Comparing these algorithms on 6 datasets by different validity indices, we get:

i. Classification accuracy: From the results in Table 3, TS3MFCM gets the best results on 5 datasets (Dermatology, Ecoli, Ionosphere, Vertebral, Wdbc). Clustering accuracy of CS3FCM is the best one on Glass. On this dataset, classification accuracy of TS3MFCM is a bit lower than that of CS3FCM (0.54 vs 0.57).

ii. Clustering quality: As showed in Table 3, TS3MFCM is the best model in term of DB index on all datasets. Moreover, the values of DB obtained by CS3FCM are much higher than the values obtained by TS3MFCM (about 2.88 times higher on average). Apart from that, the derivation of computation when applying CS3FCM is also much higher than applying TS3MFCM (about 9.82 times higher on average).

iii. Computational time: As showed in Table 3, CS3FCM is better than TS3MFCM in time consuming. This is caused by the calculation step of m_i and other extra works in TS3MFCM comparing with CS3FCM. However, the total of runtime on all six datasets by applying TS3MFCM is a bit higher than applying CS3FCM (only about 1.27 seconds higher).

On overall, TS3MFCM gets better performance than CS3FCM in term of clustering accuracy and clustering quality. In run time, TS3MFCM takes a bit longer than CS3FCM.

3.2.2. Evaluation on airbus ship detection challenge dataset

The data image belongs to the RGB color system with standard size of 768 × 768 pixels. The original image is converted to 201 × 201 pixel image using the $InterArea$ interpolation

supported in the OpenCV image processing library. From the Airbus Ship Detection Challenge Dataset [24], we use the *TrainShipSegmentations.csv* file to locate pixels containing ships. Using a 3×3 sliding window to scan the surface of the image, the obtained results are used to synthesize the result of attributes in images. Then, the number of attributes is also reduced by converting the RGB to a grayscale image. Based on the remaining attributes, our program runs 20 times for each image and initializes the label for 20% of random pixels, the other pixels are unlabeled. In labeled pixels, we run the experiments with the amount of incorrect label as 0%, 5%, 10%, 15%, 20%, 25%, 30%, respectively. To illustrate the performance of TS3MFCM and CS3FCM visually, the results of running these algorithms on four satellite ship images are given as in Figures 3- 6 below.

Figure 3: Clustering results of image **1**: a) The original image; b) By applying TS3MFCM; c) By applying CS3FCM

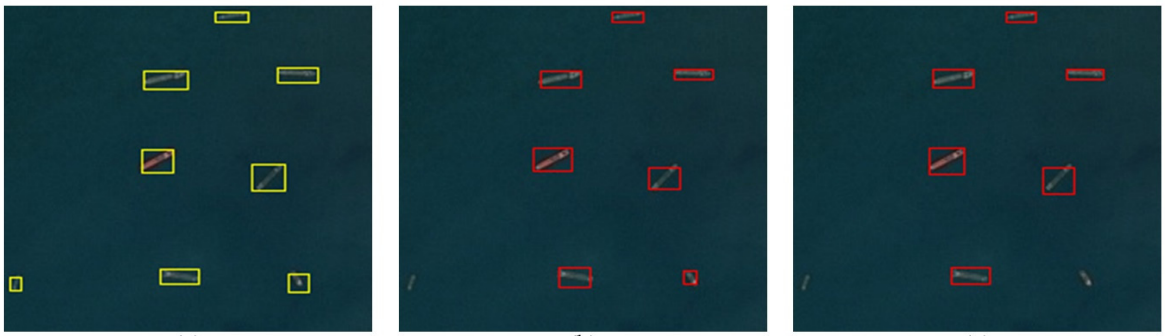
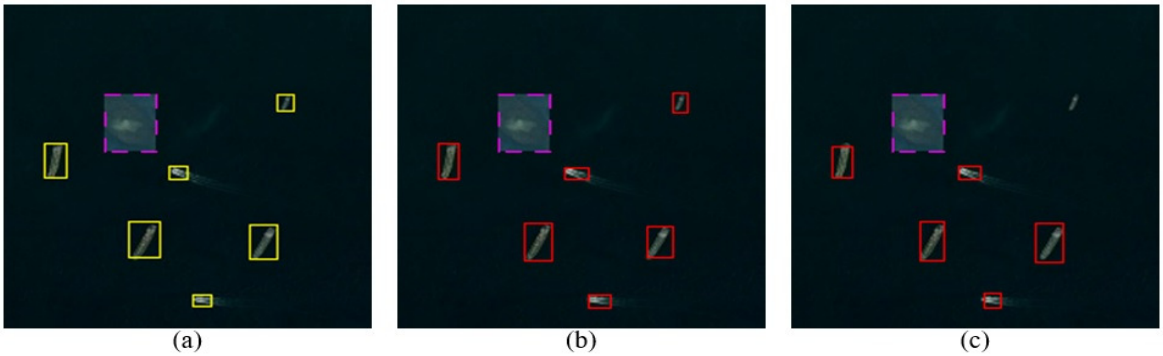


Figure 4: Clustering results of image **2**: a) The original image; b) By applying TS3MFCM; c) By applying CS3FCM

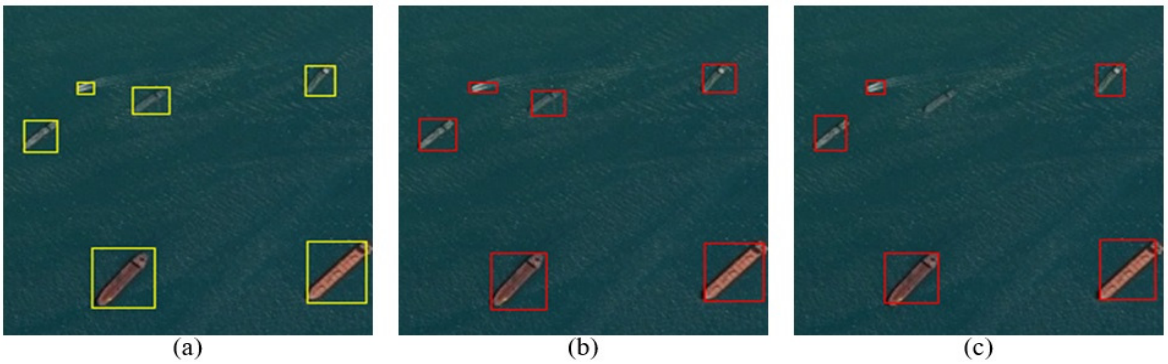
In the original images, the ships are surrounded by solid yellow rectangles. In the resulting images, the detected areas that are assumed to be ships will be surrounded by solid red rectangles. The islands on these three images are marked by purple dashed rectangles.

In Figure 3, the original image (Figure 3.a) includes 5 ships and 2 islands. The clustered image obtained by applying TS3MFCM (Figure 3.b) identifies 5 ships in which 4 ships are correctly detected. The other one is mistakenly identified as ship while that area is island. In clustered image obtained by CS3FCM (Figure 3.c), there are 6 ships are detected. However, there



(a) (b) (c)

Figure 5: Clustering results of image **3**: a) The original image; b) By applying TS3MFCM; c) By applying CS3FCM



(a) (b) (c)

Figure 6: Clustering results of image **4**: a) The original image; b) By applying TS3MFCM; c) By applying CS3FCM

are also 4 correct ships. The other two are misidentified from islands. Thus, TS3MFCM is more accurate than CS3FCM in ship and island detection from image 1.

As given in Figure 4, the original image (Figure 4.a) includes 8 ships. TS3MFCM is detected correctly 7 out of 8 ships (Figure 4.b) while CS3FCM detects correctly 6 out of 8 ships (Figure 4.c). The results of object detection by applying TS3MFCM on Image 3, 4 are also better than applying CS3FCM. To have an overall evaluation, the number of objects that are detected correctly or incorrectly by using TS3MFCM and CS3FCM on four selected images is given in Table 4.

The results on each image and the synthetic table show that clustering results of TS3MFCM is better than CS3FCM in detecting ships and islands from satellite images.

4. CONCLUSIONS

This paper proposed an improvement algorithm for the data partition with confidence problem using semi-supervised clustering and multiple fuzzifiers named as TS3MFCM. The proposed method includes 3 main steps mentioned in Section 2. The proposed method is implemented and experimentally compared to the CS3FCM in clustering accuracy, clustering quality and computational time. The findings of this research can be stated as:

Table 4. Comparison of detecting results of TS3MFCM and CS3FCM on four images

Images	No. of ships on image	No. of island on image	Object detecting results					
			No. of correct detected ships		No. of correct detected islands		No. of islands is misclassified as ships	
			TS3MFCM	CS3FCM	TS3MFCM	CS3FCM	TS3MFCM	CS3FCM
Image 1	5	2	4	4	1	0	1	2
Image 2	8	0	7	6	0	0	0	0
Image 3	6	1	6	5	1	1	0	0
Image 4	6	0	6	5	0	0	0	0

• Propose a three step model (TS3MFCM) in order to partition objects from a dataset with confidence and to deal with noise/outlier data.

• Introduce a modification of FCM using both labeled and unlabeled data. This modified FCM method supplies the evaluation on impact of labeled data.

• Present the process of using semi-supervised clustering method with different fuzzifier for each data element.

• Implement the proposed model and CS3FCM on 6 different datasets in which some datasets contain outlier data and noise data as well. The comparison of TS3MFCM and CS3FCM is performed by using validity indices and by visual results on satellite images. The visually comparison is also given by the applying two these models on four satellite images from Airbus Ship Detection Challenge DataSet to deal with object detecting problem.

The advantages of the proposed algorithm can be seen as the capability to remove or reduce the noisy data elements and the higher performance in term of clustering accuracy and clustering quality. The higher accurate in ship and island detection of TS3MFCM is also given. However, there are still some limitations of our method such as high computational time and needing many parameters. In further studies, we will develop a new algorithm to remedy these disadvantages.

ACKNOWLEDGMENT

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.05-2020.11.

REFERENCES

[1] J.C. Bezdek, *Pattern recognition with fuzzy objective function algorithms. Kluwer Academic Publishers*, 1981.

[2] Salem Saleh Al-amri, N.V. Kalyankar, and S.D. Khamitkar, "Image segmentation by using thershod techniques," *Journal of Computing*, vol. 2, no. 5, 2010, pp. 83–86.

[3] Bezdek, James C., Robert Ehrlich, and William Full, "FCM: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2–3, 1984, pp. 191–203. https://doi.org/10.1016/0098-3004(84)90020-7

[4] T. C. Havens, J. C. Bezdek, C. Leckie, L. O. Hall and M. Palaniswami, "Fuzzy c-Means Algorithms for Very Large Data," in *IEEE Transactions on Fuzzy Systems,* vol. 20, no. 6, pp. 1130–1146, Dec. 2012. Doi: 10.1109/TFUZZ.2012.2201485

[5] Seresht, N. G., Lourenzutti, R., & Fayek, A. R. (2020). A fuzzy clustering algorithm for developing predictive models in construction applications. *Applied Soft Computing*,96, 106679.

[6] H. Lu, S. Liu, H. Wei, and J. Tu, "Multi-kernel fuzzy clustering based on auto-encoder for fMRI functional network," *Expert Systems with Applications*, vol. 159, November 2020, https://doi.org/10.1016/j.eswa.2020.113513

[7] Q. T. Bui, B. Vo, V. Snasel, W. Pedrycz, T. P. Hong, N. T. Nguyen, and M. Y. Chen, "SFCM: A fuzzy clustering algorithm of extracting the shape information of data," in *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 1, pp. 75–89, Jan. 2021, Doi: 10.1109/TFUZZ.2020.3014662

[8] H. Li, and M. Wei, "Fuzzy clustering based on feature weights for multivariate time series," *Knowledge-Based Systems*, vol. 197, 7 June 2020, 105907, https://doi.org/10.1016/j.knosys.2020.105907

[9] W. Pedrycz and J. Waletzky, "Fuzzy clustering with partial supervision," in *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* vol. 27, no. 5, pp. 787-795, Oct. 1997, Doi: 10.1109/3477.623232

[10] S. Kundu, U. Maulik, and A. Mukhopadhyay, "A game theory-based approach to fuzzy clustering for pixel classification in remote sensing imagery," *Soft Comput*, vol. 25, 5121–5129, 2021, https://doi.org/10.1007/s00500-020-05514-2

[11] F. Salehi, M. R. Keyvanpour, and A. Sharifi, "SMKFC-ER: Semi-supervised multiple kernel fuzzy clustering based on entropy and relative entropy," *Information Sciences*, vol. 547, 667–688, 2021, https://doi.org/10.1016/j.ins.2020.08.094

[12] J. Xiong, X. Liu, X. Zhu, H. Zhu, H. Li and Q. Zhang, "Semi-supervised fuzzy c-means clustering optimized by simulated annealing and genetic algorithm for fault diagnosis of bearings," *IEEE Access*, vol. 8, pp. 181976-181987, 2020, Doi: 10.1109/ACCESS.2020.3021720

[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, May 2019, https://doi.org/10.1016/j.engappai.2019.02.007

[14] S.D. Mai, and L.T. Ngo, "Multiple kernel approach to semi-supervised fuzzy clustering algorithm for land-cover classification," *Engineering Applications of Artificial Intelligence,* vol. 68, pp. 205-213, February 2018, https://doi.org/10.1016/j.engappai.2017.11.007

[15] O. Komori, S. Eguchi, "A unified formulation of k-Means, fuzzy c-Means and Gaussian mixture model by the Kolmogorov–Nagumo average," *Entropy*, vol. 23, no. 5, 2021, https://doi.org/10.3390/e23050518

[16] L.H. Son, T.M. Tuan, "Dental segmentation from X-ray images using semi-supervised fuzzy clustering with spatial constraints," *Engineering Applications of Artificial Intelligence,* vol. 59, pp. 186-195, March 2017, https://doi.org/10.1016/j.engappai.2017.01.003

[17] B. Li, X. Xie, X. Wei, and W. Tang, "Ship detection and classification from optical remote sensing images: A survey," *Chinese Journal of Aeronautics*, vol. 34, no. 3, pp. 145-163, March 2021, https://doi.org/10.1016/j.cja.2020.09.022

[18] G. Casalino, G. Castellano, C. Mencar, "Data stream classification by dynamic incremental semi-supervised fuzzy clustering," *International Journal on Artificial Intelligence Tools,* vol. 28, no. 8, 2019, https://doi.org/10.1142/S0218213019600091

[19] T. D. Khang, N. D. Vuong, M. K. Tran, and M. Fowler, "Fuzzy C-means clustering algorithm with multiple fuzzifier," *Algorithms*, vol. 13, no. 7, 2020, https://doi.org/10.3390/a13070158

[20] H. Gan, "Safe semi-supervised fuzzy c -Means clustering," *IEEE Access*, vol. 7, pp. 95659-95664, 2019, Doi: 10.1109/ACCESS.2019.2929307

[21] 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, https://doi.org/10.1016/j.eswa.2017.12.046

[22] L. Lovász, M. D. Plummer, *Matching theory*, vol. 367, Ams Chelsea Publishing, 2009.

[23] Outlier Detection DataSets (2021). Data. Online: http://odds.cs.stonybrook.edu/

[24] Satellite Image DataSets of Ships (2018). Data. Online: https://www.kaggle.com/c/airbus-ship-detection/

[25] C. Hwang and F. C. Rhee, "Uncertain fuzzy clustering: Interval type-2 fuzzy approach to c-Means," in *IEEE Transactions on Fuzzy Systems*, vol. 15, no. 1, pp. 107-120, Feb. 2007, Doi: 10.1109/TFUZZ.2006.889763

[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, https://doi.org/10.1002/sam.10080

Received on November 10, 2021
Accepted on February 05, 2022