General Semi-supervised Possibilistic Fuzzy c-Means clustering for Land-cover Classification

Dinh Sinh Mai
Institute of Techniques for Special Engineering
Le Quy Don Technical University
Hanoi, Vietnam
maidinhsinh@gmail.com

Long Thanh Ngo
Institute of Simulation Technology
Le Quy Don Technical University
Hanoi, Vietnam
ngotlong@gmail.com

***Abstract*—Satellite images with the advantage of wide coverage, short update times can help to establish land-cover maps quickly and efficiently. However, due to the influence of natural conditions, satellite images often contain noise, outliers, the boundary of the objects on the image is unclear and this makes it difficult for many clustering algorithms. The possibilistic fuzzy c-means clustering (PFCM) algorithm has advantages of both fuzzy c-means clustering (FCM) and possibilistic c-means clustering (PCM) algorithms due to the simultaneous use of both fuzzy and function functions, but it also has limitations such as sensitivity with noise and outliers. The paper proposes a general semi-supervised possibilistic fuzzy c-means clustering (GSPFCM) algorithm to improve the clustering quality of PFCM. Our proposed method can solve problems that labeled data has very little compared to unlabeled data. Results of land-cover classification using satellite images (Landsat-7 ETM+, Sentinel-2A) show that the proposed method can significantly improve the accuracy of classification results when compared to some previous methods.**

***Index Terms*—Semi-supervised, satellite image, fuzzy clustering, possibilistic.**

I. Introduction

In clustering, there are two widely used methods: hard clustering and soft (fuzzy) clustering. In hard clustering, data samples can only belong to (probability 1) or not (probability equal to 0) to a single cluster, this is difficult to handle on data where patterns can simultaneously belong to many different clusters.

In 1965, Zadeh was the first to introduce the fuzzy set [1], allowing data patterns to simultaneously belong to many different clusters. The algorithm is widely used as the basic theory for fuzzy clustering problems is the fuzzy c-means clustering (FCM) algorithm, the original idea was introduced by Dunn [2], then completed and introduced by Bezdek [3] in 1984. According to the FCM algorithm, the membership function values will be calculated based on the distance between the patterns to the cluster centers, high values indicate that the data sample is closer to the cluster center. There are many ways to determine the distance between the data pattern and the cluster centers, which is most commonly used as the Euclidean distance. This distance is good in cases where the clusters are spherical, but not good in the case of complex shapes, overlapping data. Furthermore, this algorithm is also showed to be sensitive to noise and unusual elements [3].

Due to the impact of urbanization, the land-cover area is constantly changing, the establishment of a land-cover map by traditional methods is increasingly difficult. Satellite image data has the advantage of wide coverage, fast updating time, but they also have many disadvantages such as being affected by weather conditions, image data often contains noise, boundaries between objects are often unclear. Due to its inherent complexity, the problem of satellite image analysis is always a challenging task. Although the original FCM algorithm has been widely used in the past, it has shown many disadvantages such as sensitivity to noise, outliers and does not handle well on data sets with high uncertainty like satellite image data [5].

Similar to the approach based on fuzzy sets, the possibilistic approach introduced by Krishnapuram and J. Keller [6] allowing the determination of possibilistic partitions based on possibilistic membership. The possibilistic membership value is determined by Euclidean distance, the small value represents the large possibilistic membership grade. This method has the disadvantage that it is difficult to separate similar clusters [7]. For improvement, Zhang et al. [8] proposed a possibilistic approach based on c-means clustering (PCM) to deal with similar clusters. However, PCM still has difficulty in selecting parameters and they are not effective for data with complex structures and shapes.

After that, in 2005, the PFCM model was proposed by Nikhil et al. [9], this is a hybrid model between FCM and PCM algorithms to deal with the disadvantages of PCM and FCM. PFCM algorithm still has the disadvantages of type-1 fuzzy set and difficulty in selecting parameters for the algorithm and sensitive with noise. An improvement of PFCM based on entropy introduced by Askari et al. [10]. This algorithm uses a combination of the c-means general entropy algorithm (ECM) and the PFCM algorithm to deal with noisy data.

The semi-supervised model is introduced by Yasunori et al. [4] only uses the fuzzy membership function (MF) constraint, which may not be suitable for the objective functions with many MFs [9]. There have been some studies to improve the unsupervised clustering algorithms based on the semi-supervised method [11], [12], [13], but these studies only use constraints with fuzzy MF or with cluster centroids [14], [15], [16]. In this study, we introduce a general semi-supervised PFCM clustering algorithm (GSPFCM) for the

978-1-7281-3003-3/19/$31.00 ©2019 IEEE

problem of satellite image land-cover classification. Unlike previous studies, all constraints on fuzzy MF U, possibilistic MF T and cluster centroid V are all built from labeled data. The contribution of the paper was to provide a generalized semi-supervised PFCM model that could be used in the case of very few labeled data or additional information.

The paper includes the following sections: Section I Introduction; Section II Background; Section III General Semi-supervised Possibilistic Fuzzy c-Means clustering; Section IV Some experimental results; Section V Conclusion and some future research directions.

II. Backgrounds

The PFCM model was introduced by Nikhil and his colleagues [9]. This is the hybrid algorithm between FCM and PCM algorithms, the advantage of this algorithm is to use fuzzy MF and possibilistic MF simultaneously to describe data. The PFCM model is the constrained optimization problem:

$$\min\{J_{m,\eta}(U,T,V,X,\gamma) = \sum_{i=1}^{c}\sum_{k=1}^{n}(a\mu_{ik}^{m} + b\tau_{ik}^{\eta})d_{ik}^{2} + \sum_{i=1}^{c}\gamma_i\sum_{k=1}^{n}(1-\tau_{ik})^{\eta}\} \quad (1)$$

Where $X = \{\mathrm{x}_k,\mathrm{x}_k \in \mathrm{R}^M, \mathrm{k} = 1,...,\mathrm{n}\}$, $U = [\mu_{ik}]_{cxn}$ is a fuzzy MF matrix, which contains the fuzzy membership degree, $T = [\tau_{ik}]_{cxn}$ is a typicality MF matrix, which contains the possibilistic membership degree, $V = (v_1, v_2, ..., v_c)$ is a vector of cluster centers, m is the weighting exponent for fuzzy MF matrix and η is the weighting exponent for typicality MF matrix. $\gamma_i > 0$ are constants given by the user. Instead of having only one function type such as FCM and PCM, the PFCM algorithm has two types of MFs that are the fuzzy MF and the possibilistic MF.

Subject to the constraints:

$$m,\eta > 1; a,b > 0; 0 \le \mu_{ik},\tau_{ik} \le 1; \sum_{i=1}^{c}\mu_{ik} = 1; \sum_{k=1}^{n}\tau_{ik} = 1; 1 \le i \le c; 1 \le k \le n \quad (2)$$

The objective function $J_{m,\eta}(U,T,V,X)$ reaches the smallest value with the constraints 2 when and only if:

$$v_i = \frac{\sum_{k=1}^{n}(a\mu_{ik}^{m} + b\tau_{ik}^{\eta})x_i}{\sum_{k=1}^{n}(a\mu_{ik}^{m} + b\tau_{ik}^{\eta})} \quad (3)$$

$$\mu_{ik} = 1/\sum_{j=1}^{c}\left(d_{ik}^{2}/d_{jk}^{2}\right)^{2/(m-1)} \quad (4)$$

$$\tau_{ik} = 1/\left(1 + \left(bd_{ik}^{2}/\gamma_i\right)^{1/(\eta-1)}\right) \quad (5)$$

In which, with equations 3, 4, 5 are achieved by using the Lagrange operator to minimize the objective function 2.

Details of the implementation steps of the PFCM algorithm described below:

Algorithm 1: PFCM algorithm

Input: A dataset $X = \{\mathrm{x}_k,\mathrm{x}_k \in \mathrm{R}^M, \mathrm{k} = 1,...,\mathrm{n}\}$, the number of clusters C $(1 < C < n)$, fuzzifiers m, η, $T_{\max}, t = 0$.

Output: The membership matrix U, T and the cluster centroid V.

Step 1: Initialize the cluster centroid $V^{(0)} = [v_i^{(0)}], V^{(0)} \in R^{MxC}$ by choosing randomly from the input dataset X.

Step 2: Compute $U^{(0)}$ by using the equation 4 and $T^{(0)}$ by using the equation 5.

Step 3: Loop

3.1 t=t+1

3.2 Compute $V^{(t)} = [v_1^{(t)}, v_2^{(t)}, ..., v_C^{(t)}]$ by using equation 3.

3.3 Compute $U^{(t)} = [\mu_{ik}^{(t)}]$ by using equation 4.

3.4 Compute $T^{(t)} = [\tau_{ik}^{(t)}]$ by using equation 5.

3.5 Check if $\max(\left\|U^{(t+1)} - U^{(t)}\right\| + \left\|T^{(t+1)} - T^{(t)}\right\|) \le \varepsilon$ or $t > T_{max}$ then stop else go to Step 3.

Defuzzification: Assign data x_k to the i^{th} cluster if $u_{ik} \ge u_{jk}, j = 1, ..., C; j \ne C$.

Computational complexity: The PFCM algorithm will execute a conditional loop, when either of the conditions $\max(\left\|U^{(t+1)} - U^{(t)}\right\| + \left\|T^{(t+1)} - T^{(t)}\right\|) \le \varepsilon$ or $t > T_{max}$ comes first, the algorithm will stop and give the classification result. Each loop will calculate V, U and T according to equations 3, 4 and 5. The algorithm stops at the t^{th} loop, the computational complexity of the algorithm will be **O(3tnMC)**.

III. General Semi-supervised PFCM clustering

A. Proposed method

In this part, we present a general semi-supervised algorithm based on PFCM algorithm. Considered data $X = \{\mathrm{x}_k,\mathrm{x}_k \in \mathrm{R}^M, \mathrm{k} = 1,...,\mathrm{n}\}$, with $\mathrm{X} = \mathrm{X}_1 \cup \mathrm{X}_2$, $X_1 = [x_1^*, x_2^*, ..., x_L^*]$ is a labeled data set and $X_2 = [x_{L+1}, x_{L+2}, ..., x_n]$ is an unlabeled data set ($|\mathrm{X}_1| << |\mathrm{X}_2|$).

From the labeled data set, the centroid constraints $V^* = [v_1^*, v_2^*, ..., v_c^*]$ will be calculated by averaging, C is the number of clusters.

The constraint of fuzzy membership function $U^* = [\mu_{ik}^*]$ calculated by equation:

$$\mu_{ik}^* = 1/\sum_{z=1}^{c}\left(\frac{x_k - v_i^*}{x_k - v_z^*}\right)^{2/(m-1)} \quad (6)$$

In equation 5, the T value is a constant defined by the user, but in the research [7], Krishnapuram and Keller also suggest using fuzzy MF as a good way to initialize the parameter T according to the following formula:

$$\gamma_i = K\sum_{k=1}^{n}(\mu_{ik})^{\eta}d_{ik}^{2}/\sum_{k=1}^{n}(\mu_{ik})^{\eta} \quad (7)$$

Where μ_{ik} is the fuzzy MF value from the results of the equation 6, K is a user-defined constant (usually selected by 1). The constraint of possibilistic membership function $T^{(*)} = [\tau_{ik}^{(*)}]$ calculated by equation 7 and 5.

From 3 constraints on fuzzy membership function $U^* = [\mu_{ik}^*]$, possibilistic membership function $T^{(*)} = [\tau_{ik}^{(*)}]$ and

cluster centroids $V^* = [v_1^*, v_2^*, ..., v_c^*]$, we propose a new objective function $J_{m,\eta}(U, T, V, X, \gamma)$ as follows:

$$J_{m,\eta} = \sum_{i=1}^{c}\sum_{k=1}^{n}(a\|\mu_{ik} - \mu_{ik}^*\|^m + b\|\tau_{ik} - \tau_{ik}^*\|^\eta)(\|v_i - x_k\|^2 + \delta\|v_i - v_i^*\|^2) + \sum_{i=1}^{c}\gamma_i\sum_{k=1}^{n}(1-\tau_{ik})^\eta \tag{8}$$

With to the constraints:

$$0 \le \mu_{ik}, \tau_{ik} \le 1; \sum_{i=1}^{c}\mu_{ik} = 1; \sum_{k=1}^{n}\tau_{ik} = 1; 1 \le i \le c; 1 \le k \le n \tag{9}$$

Parameters a, b and δ are user-defined constants, representing the importance of constraints, $m, \eta > 1; a, b > 0; \delta \ge 0$. $\delta = 0$ when v_i^* does not exist.

Set $D_{ik}^2 = \|v_i - x_k\|^2 + \delta\|v_i - v_i^*\|^2$.

The GSPFCM algorithm is stated as follows: $X = \{\mathrm{x_k}, \mathrm{x_k} \in \mathrm{R}^M, \mathrm{k} = 1,...,\mathrm{n}\}$, X contains at least c distinct data points. With the constraint 9 then $J_{m,\eta}(U, T, V, X, \gamma)$ may minimize if only:

$$\mu_{ik} = \mu_{ik}^* + \frac{(1 - \sum_{i=1}^{c}\mu_{ik}^*)\left[1/D_{ik}^2\right]^{1/(m-1)}}{\sum_{i=1}^{c}\left[1/D_{ik}^2\right]^{1/(m-1)}} \tag{10}$$

$$\tau_{ik} = \begin{cases} \left(\tau_{ik}^* + \left[\frac{\gamma_i}{bD_{ik}^2}\right]^{\frac{1}{\eta-1}}\right) / \left(1 + \left[\frac{\gamma_i}{bD_{ik}^2}\right]^{\frac{1}{\eta-1}}\right) & \tau_{ik} \ge \tau_{ik}^* \\ \left(\tau_{ik}^* - \left[\frac{\gamma_i}{bD_{ik}^2}\right]^{\frac{1}{\eta-1}}\right) / \left(1 - \left[\frac{\gamma_i}{bD_{ik}^2}\right]^{\frac{1}{\eta-1}}\right) & else \end{cases} \tag{11}$$

$$v_i = \frac{\sum_{k=1}^{n}(a\|\mu_{ik} - \mu_{ik}^*\|^m + b\|\tau_{ik} - \tau_{ik}^*\|^\eta)(x_k + v_i^*)}{\sum_{k=1}^{n}(a\|\mu_{ik} - \mu_{ik}^*\|^m + b\|\tau_{ik} - \tau_{ik}^*\|^\eta)(1+\delta)} \tag{12}$$

Equation 10 can be achieved by using the Lagrange multiplier with fixed T and V by minimum problem: $\min\left\{\sum_{i=1}^{c}\sum_{k=1}^{n}(a\|\mu_{ik} - \mu_{ik}^*\|^m)(\|v_i - x_k\|^2 + \delta\|v_i - v_i^*\|^2)\right\}$ It can be seen that equation 10 is independent of the constant a and depends only on v_i and v_i^*. When $\mu_{ik}^* = 0$ (μ_{ik}^* does not exist or not use), if considering the distance D_{ik} is similar to the distance d_{ik} then equation 10 is similar the fuzzy membership in FCM algorithm.

Equation 11 is achieved by handling the minimum problem for the objective function 8, with V and U fixed by minimum problem: $\min\left\{(a\|\mu_{ik} - \mu_{ik}^*\|^m + b\|\tau_{ik} - \tau_{ik}^*\|^\eta)D_{ik}^2 + \gamma_i(1-\tau_{ik})^\eta\right\}$. When $\tau_{ik}^* = 0$ (τ_{ik}^* does not exist or not use), if considering the distance D_{ik} is similar to the distance d_{ik} then equation 11 is similar the possibilistic membership in PCM algorithm.

Similarly, equation 12 is achieved by minimizing the following objective function with fixed U and T: $\min\left\{\sum_{k=1}^{n}(\|\mu_{ik} - \mu_{ik}^*\|^m + \|\tau_{ik} - \tau_{ik}^*\|^\eta)D_{ik}^2\right\}$. If v_i^* is not used or not exist then $\delta = 0$. In equation 12, if additional information (v_i^*, τ_{ik}^*, μ_{ik}^*) is not used, they will become the equation 3 in PFCM.

Without reducing the generality, the additional information μ_{ik}^*, τ_{ik}^*, v_i^* can be achieved by different methods. May be from labeled data, expert experience or results from other methods. The calculation of μ_{ik}^*, τ_{ik}^*, v_i^* in this study is only one of them.

Algorithm 2: GSPFCM algorithm

Input: A dataset $X = X_1 \cup X_2$, $X_1 = [x_1^*, x_2^*, ..., x_L^*]$, $X_2 = [x_{L+1}, x_{L+2}, ..., x_n]$ ($|\mathrm{X_1}| << |\mathrm{X_2}|$), the number of clusters C ($1 < C < n$), fuzzifiers $m, \eta > 1$, $T_{\max}, t = 0$, $a, b > 0; \delta \ge 0$.

Output: The membership matrix U, T and the centroid matrix V.

Step 1: Compute $V^{(*)} = [v_i^{(*)}], V^{(*)} \in R^{MxC}$ from X_1.

Step 2: Compute $U^{(*)} = [\mu_{ik}^{(*)}]$ by using equation 6.

Step 3: Compute $T^{(*)} = [\tau_{ik}^{(*)}]$ by using equation 7 and equation 5.

Step 4: Initialize the centroid matrix $V^{(0)}$ and fuzzy membership function $U^{(0)}$ by running the FCM algorithm on dataset X.

Step 5: Compute $T^{(0)}$ by using equation 7 and equation 5.

Step 6: Loop

6.1 t=t+1

6.2 Compute $V^{(t)} = [v_1^{(t)}, v_2^{(t)}, ..., v_C^{(t)}]$ by using equation 12.

6.3 Compute $U^{(t)} = [\mu_{ik}^{(t)}]$ by using equation 10.

6.4 Compute $T^{(t)} = [\tau_{ik}^{(t)}]$ by using equation 7 and 11.

6.5 Check if $\max(\left\|U^{(t+1)} - U^{(t)}\right\| + \left\|T^{(t+1)} - T^{(t)}\right\|) \le \varepsilon$ or $t > T_{max}$ then stop else go to Step 6.1.

Defuzzification: Assign data x_k to the i^{th} cluster if $u_{ik} \ge u_{jk}, j = 1, ..., C; j \ne C$.

Computational complexity: The GSPFCM algorithm will execute a conditional loop, when either of the conditions $\max(\left\|U^{(t+1)} - U^{(t)}\right\| + \left\|T^{(t+1)} - T^{(t)}\right\|) \le \varepsilon$ or $t > T_{max}$ comes first, the algorithm will stop and give the classification result. Each loop will calculate V, U and T according to equations 12, 10, 7 and 11. The algorithm stops at the t^{th} loop, the computational complexity of the algorithm will be **O(4tnMC)**. When n is large, the computational complexity of GSPFCM and PFCM algorithm is the same.

B. Evaluation methods

To compare classification results, we use labeled data to check the correct classification rate and wrong classification rate.

On the other hand, the clustering results have been evaluated by some validity indexes including Bezdeks partition coefficient index (PC-I) [17], Classification Entropy index (CE-I) [18], Xie-Beni index (XB-I) [19], and τ index ($\tau - I$) was introduced to assess the degree of characteristic separation between pixels and cluster centrois, Mean Squared Error index (MSE) [20]. Large values with index PC-I are good clustering results, while small values with indexes CE-I, XB-I, $\tau - I$ and MSE are good clustering results.

However, the GSPFCM algorithm has two membership functions, so we change some formulas to calculate indexes. Details of the formulas are as follows:

- Partition Coefficient index:

$$PC = \frac{1}{n}\sum_{i=1}^{C}\sum_{k=1}^{n}(\mu_{ik}^2 + \tau_{ik}^2) \tag{13}$$

- Classification Entropy index

$$CE = -\frac{1}{n}\sum_{i=1}^{C}\sum_{k=1}^{n}(\mu_{ik}\log\mu_{ik} + \tau_{ik}\log\tau_{ik}) \tag{14}$$

- Xie and Benis index:

$$XB = \frac{1}{n}\sum_{i=1}^{C}\sum_{k=1}^{n}\mu_{ik}^m D_{ik}^2 / \min_{i,j=1,\ldots,C;i\neq j}\|v_i - v_j\|^2 \tag{15}$$

- The index τ is calculated as follows:

$$\tau = \frac{1}{n}\sum_{i=1}^{C}\sum_{k=1}^{n}\tau_{ik}^{\eta} D_{ik}^2 / \min_{i=1,\ldots,C;\forall x_k\notin v_i}\|v_i - x_k\|^2 \tag{16}$$

- MSE index:

$$MSE = \frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n}(x_{ik} - v_i)^2 \tag{17}$$

In which $X = \{x_i\} = \{x_{i1}, x_{i2}, ..., x_{in}\}$ and $V = \{v_i\} = \{v_1, v_2, ..., v_c\}$ respectively the initial pixels and the centroid of the clusters. The small MSE index represents clustering results as well.

IV. Experiment

A. Initialize parameters

We selected two datasets at locations with different characteristics including (city, delta and mountain forest) on different images for testing. Multi-spectral satellite images used include Sentinel-2A and Landsat-7 ETM+. The experimental data is clustered into 6 classes with colors described as follows: 1. Rivers, ponds, lakes ; 2. Rocks, bare soil ; 3. Fields, grass ; 4. Planted forests ; 5. Perennial forest ; 6. Jungles forest . The labeled data is taken directly from the pixels according 6 landcover classes on the satellite image.

Tested on algorithms PFCM, SPFCM-W [14], SPFCM-SS [15], GSPFCM. Parameters of PFCM, SPFCM-SS algorithms and SPFCM-W algorithms are taken from the original papers. $m = \eta = 2$, $a = b = \delta = 1$, select K = 1 to calculate the value γ_i, $\varepsilon = 10^{-6}$, $T_{max} = 1000$

B. Experiment 1: Landsat-7 ETM+ imagery

Test data is the Landsat-7 ETM+ satellite image of Phan Thiet city region, Binh Thuan province on February 12, 2010, $107^0 54'6.0126"E$, $11^0 02'45.1138"N$ and $108^0 15'55.3798"E$, $10^0 53'47.5691"N$ with 7 spectrum bands (see Figure 1). In the experiment, we only use 6 spectrum bands with $30m$ resolution. The number of pixels is 1.048.576 pixels and the area is $94371.84ha$. Labeling data is taken about 1% and divided equally among classes.

Fig. 1. Landsat-7 ETM+ imagery: The RGB image of Phan Thiet city of Binh Thuan province

Figure 2 is the result of the land-cover classification according to algorithms PFCM, SPFCM-W, SPFCM-SS and GSPFCM, respectively. It can be seen that the central of Phan Thiet city, the classification results in figure 2.c and figure 2.d are better than figure 2.a and figure 2.b. Especially between class 1 and class 2, the SPFCM-SS and GSPFCM algorithms have better separation capabilities than the other two algorithms.

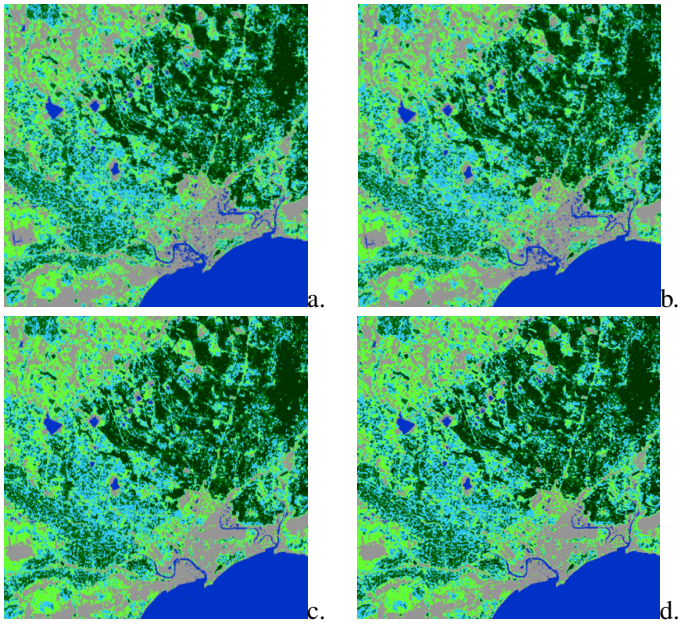

Fig. 2. Result of landcover classification of Phan Thiet region: a) PFCM; b) SPFCM-W; c) SPFCM-SS; d) GSPFCM

The classification results also shown in Table I, which is the correct classification rate achieved by labeled data. Accuracy

of class 1 and class 2 is more than 95% with SPFCM-SS and GSPFCM algorithms. While the SPFCM-SS algorithm gives the best results in class 4 with 92.81%, the GSPFCM algorithm gives the best results in all the remaining classes. The highest correct classification rate is 95.01% with GSPFCM algorithm, followed by 93.16%, 92.20%, 85.50% with algorithms SPFCM-SS, SPFCM-W and PFCM, respectively.

TABLE I
CORRECT CLASSIFICATION RATE OBTAINED ACCORDING TO THE LABELED DATA FOR PHAN THIET REGION

No. Class	Algorithm			
	PFCM	***SPFCM-W***	***SPFCM-SS***	***GSPFCM***
1	89.32 %	94.82 %	95.89 %	**97.38 %**
2	87.49 %	92.96 %	95.28 %	**96.12 %**
3	85.87 %	93.11 %	93.75 %	**96.29 %**
4	86.83 %	89.99 %	**92.81 %**	92.69 %
5	82.09 %	91.87 %	90.04 %	**93.64 %**
6	81.38 %	90.45 %	91.21 %	**93.92 %**
Total	85.50 %	92.20 %	93.16 %	**95.01 %**

Table II is the value of cluster quality indicators, it can be seen that the GSPFCM algorithm gives the best results in the indexes CE, XB, τ, MSE with values of 0.2987; 1,0662; 0.0389; 10.3542, respectively. while the SPFCM-SS algorithm gives the best results at the PC index with a value of 0.6698. According to Table II, the PFCM algorithm gives the worst classification results.

TABLE II
VALIDY INDICES OBTAINED FOR PHAN THIET REGION

Index Algorithm	Validy indices				
	PC-I	***CE-I***	***XB-I***	$\tau - I$	***MSE***
PFCM	0.4898	0.4562	1.8983	0.0789	20.7634
SPFCM-W	0.6547	0.3876	1.4637	0.0563	16.4678
SPFCM-SS	**0.6698**	0.3872	1.3988	0.0478	14.7874
GSPFCM	0.6621	0.2987	**1.0662**	**0.0389**	**10.3542**

C. Experiment 2: Sentinel-2A imagery

Test data is Sentinel-2A image of Tam Dao mountain region, Vinh Phuc province on December 20, 2017 with 12 spectrum bands (see Figure 3). In the experiment, we only use 4 spectrum bands with $10m$ resolution. The number of pixels is 491.401 pixels and the area is $4914.01ha$. Labeling data is taken about 1% and divided equally among classes.

Figure 4 shows the classification results of Tam Dao region. In Figure 3, the southwestern region of Tam Dao has a light colored region of urban land, The classification results by PFCM, SPFCM-W, SPFCM-SS algorithms at this location are not good, they are mixed by class 1. While the GSPFCM algorithm gives better classification results.

The classification results also shown in Table III, which is the correct classification rate achieved by labeled data. The highest correct classification rate is 97.26% with GSPFCM algorithm, followed by 94.29%, 94.04%, 89.11% with algorithms SPFCM-SS, SPFCM-W and PFCM, respectively.

Table IV is the value of cluster quality indicators, it can be seen that the GSPFCM algorithm gives the best results

Fig. 3. Sentinel-2A imagery: The RGB image of Tam Dao mountain region, Vinh Phuc province

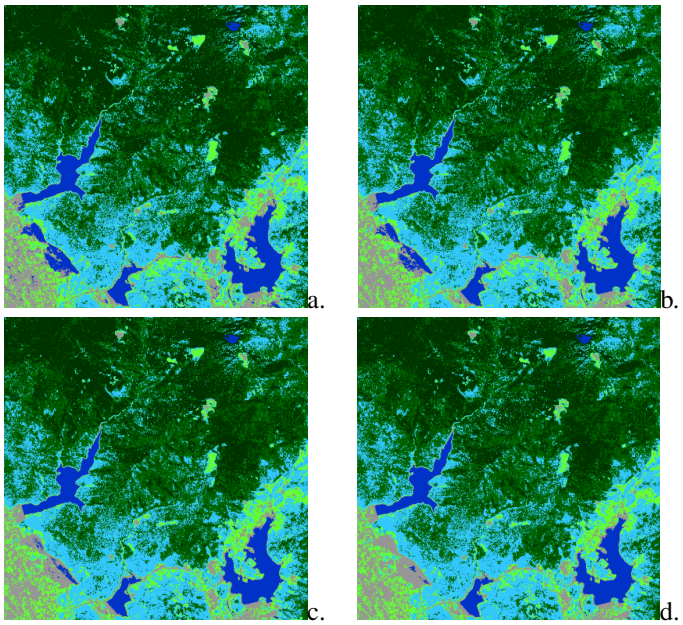

Fig. 4. Result of landcover classification of Tam Dao region: a) PFCM; b) SPFCM-W; c) SPFCM-SS; d) GSPFCM

at all indicators and the PFCM algorithm gives the worst classification results. While the SPFCM-SS algorithm gives better results than the SPFCM-W algorithm in PC, τ and MSE indices; but it is not good at CE and XB indicators.

From the above tests, it can be seen that the GSPFCM algorithm gives better results than the remaining algorithms in most indicators. The results compared with labeled data also show that the GSPFCM algorithm has higher accuracy with the rate of over 95%.

In this study, we tested on two satellite images with different

TABLE III
CORRECT CLASSIFICATION RATE OBTAINED ACCORDING TO THE LABELED DATA FOR TAM DAO REGION

No. Class	Algorithm			
	PFCM	***SPFCM-W***	***SPFCM-SS***	***GSPFCM***
1	92.43 %	96.36 %	96.83 %	**99.19 %**
2	89.78 %	94.92 %	95.61 %	**98.83 %**
3	91.67 %	93.81 %	94.96 %	**97.27 %**
4	88.39 %	92.67 %	92.79 %	**95.96 %**
5	86.45 %	93.56 %	93.85 %	**96.62 %**
6	85.92 %	92.94 %	91.68 %	**95.71 %**
Total	89.11 %	94.04 %	94.29 %	**97.26 %**

TABLE IV
VALIDY INDICES OBTAINED FOR TAM DAO REGION

Index Algorithm	Validy indices				
	PC-I	***CE-I***	***XB-I***	$\tau - I$	***MSE***
PFCM	0.6783	0.5655	1.3279	0.0687	15.6783
SPFCM-W	0.7837	0.4681	1.0983	0.0598	13.7949
SPFCM-SS	0.8832	0.4682	1.0998	0.0487	12.4527
GSPFCM	**0.8871**	**0.3873**	**0.9986**	**0.0269**	**8.3674**

resolutions, which also affected the accuracy of the classification results. The Sentinel-2A image has a resolution of $10m$, while the Landsat-7 ETM+ image is $30m$. Tables I and III show that classification from Sentinel-2A image data will give more accurate results from Landsat-7 ETM+ image at all algorithms.

However, the image resolution is also related to cost and calculation time. In the same region with high-resolution satellite images, the number of pixels will be higher than the number of pixels of low-resolution satellite images, so the calculation will be slower although the accuracy will be higher. Therefore, depending on the accuracy of each problem to select satellite image data with appropriate resolution.

V. CONCLUSION

The paper proposes the GSPFCM algorithm to generalize the semi-supervised method for the PFCM algorithm, which is an open algorithm that allows using one or more constraint parameters U^*, T^* and V^*. Our proposed method can also solve problems that labeled data has very little compared to unlabeled data. Tested on Landsat-7 ETM+ and Sentinel-2A satellite image data for land-cover classification problem shows that if GSPFCM algorithm used constraints simultaneously from labelled data for the fuzzy membership function, possibilistic membership function and cluster centroids, the classification results will be better in most cases when compared to PFCM, SPFCM-W and SPFCM-SS algorithms.

In the future, we will study to improve this algorithm based on type-2 fuzzy set, optimize parameters and calculations on GPUs.

ACKNOWLEDGMENT

This research is funded by the Newton Fund, under the NAFOSTED - UK Academies collaboration programme. According to Decision No. 04/QD-HDQL-NAFOSTED, January 7, 2019 of National Foundation for Science and Technology Development, Vietnam.

REFERENCES

[1] L.A. Zadeh. Fuzzy Sets, Information and Control, vol. 8, pp.338-353, 1965.
[2] J.C. Dunn. A fuzzy relative of the ISO-DATA process and its use in detecting compacy well-separated clusters, J. Cybernetics, vol. 3, pp. 32-57, 1973.
[3] J.C. Bezdek, E.R William, Full. FCM: The fuzzy c-meams clustering algorithm. Computer and Geoscience. 10 (23), 191-203, 1984.
[4] J. Kennedy, R. Eberhart. Particle Swarm Optimization. IEEE International Conference on Neural Networks, pp. 1942-1948, 1995.
[5] T.L Hung, D.S Mai. Classification of Remote Sensing Imagery Based on Density and Fuzzy c-Means Algorithm, International Journal of Fuzzy System Applications, vol.8 (2), pp.1-15, 2019.
[6] R. Krishnapuram, J. Keller. A possibilistic approach to clustering, IEEE Transactions on Fuzzy Systems, vol. 1, pp. 98-110, 1993.
[7] R. Krishnapuram, J. Keller. The possibilistic c-Means algorithm: Insights and recommendations, IEEE Transactions on Fuzzy Systems, vol. 4, no. 3, pp. 385-393, 1996.
[8] J.S Zhang, Y.W. Leung. Improved Possibilistic C-Means Clustering Algorithms, IEEE Trans on Fuzzy Systems, vol.12(2), pp.209-217, 2004.
[9] N. R. Pal, K. Pal, J. M. Keller, J. C. Bezdek. A Possibilistic Fuzzy c-Means Clustering Algorithm, IEEE Transactions on Fuzzy Systems, Vol. 13, No. 4, pp.517-530, 2005.
[10] S. Askari, N. Montazerin, M.H. Fazel Zarandi, E. Hakimi. Generalized entropy based possibilistic fuzzy C-means for clustering noisy data and its convergence proof, Neurocomputing, Vol.219, 5, pp.186-202, 2017.
[11] 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, 2017.
[12] D.S Mai, L.T Ngo, T.L Hung. Satellite Image Classification based Spatial-Spectral Fuzzy Clustering Algorithm. ACIIDS, pp. 505-518, 2018.
[13] D.S Mai, L.T Ngo. Semi-Supervised Fuzzy C-Means Clustering for Change Detection from Multispectral Satellite Image, FUZZ-IEEE, pp.1-8, 2015.
[14] D.S Mai, L.T Ngo. Semi-supervised method with spatial weights based Possibilistic fuzzy C-means clustering for Land-cover Classification, NICS 2018, pp.406-411, 2018.
[15] D.S Mai, L.T Ngo, T.L Hung. Advanced Semi-supervised Possibilistic Fuzzy C-means Clustering using Spatial-Spectral distance for Land-cover Classification, SMC 2018, pp.4375-4380, 2018.
[16] D.S Mai, 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, 2018.
[17] J.C. Bezdek, N. Pal. Some new indexes of cluster validity. IEEE Transactions on Systems, Man and Cybernetics, 28(3), pp.301-315. 1998.
[18] C.H. Chou, M.C, Su, E. Lai. A new cluster validity measure and its application to image compression. Pattern Anal Applic, 7(2), pp 205-220, 2004.
[19] U. Maulik, S. Bandyopadhyay. Performance Evaluation of Some Clustering Algorithms and Validity Indices, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(12), pp.1650-1654, 2002.
[20] Z. Wang, A.C Bovik. Mean squared error: love it or leave it? A new look at signal fidelity measures. IEEE signal processing magazine, pp.98-117, 2009.