Applied Soft Computing Journal 88 (2020) 106016

Contents lists available at ScienceDirect

Applied Soft Computing Journal

journal homepage: www.elsevier.com/locate/asoc

Fuzzy c-means clustering using Jeffreys-divergence based similarity measure

Ayan Seal [a,c,*], Aditya Karlekar [b], Ondrej Krejcar [c], Consuelo Gonzalo-Martin [d]

[a] PDPM Indian Institute of Information Technology, Design and Manufacturing, Jabalpur, 482005, India
[b] Hitkarini College of Engineering and Technology, Jabalpur, 482005, India
[c] Faculty of Informatics and Management, Center for Basic and Applied Research, University of Hradec Kralove, Rokitanskeho 62, Hradec Kralove, 50003, Czech Republic
[d] Center for Biomedical Technology, Universidad Politecnica de Madrid, Madrid, 28223, Spain

ARTICLE INFO

Article history:
Received 1 November 2018
Received in revised form 1 December 2019
Accepted 10 December 2019
Available online 17 December 2019

Keywords:
Jeffreys-divergence
Jeffreys-divergence based similarity measure
Fuzzy c-means
Jeffreys-fuzzy-c-means clustering

ABSTRACT

In clustering, similarity measure has been one of the major factors for discovering the natural grouping of a given dataset by identifying hidden patterns. To determine a suitable similarity measure is an open problem in clustering analysis for several years. The purpose of this study is to make known a divergence based similarity measure. The notion of the proposed similarity measure is derived from Jeffrey-divergence. Various features of the proposed similarity measure are explained. Afterwards we develop fuzzy c-means (FCM) by making use of the proposed similarity measure, which guarantees to converge to local minima. The various characteristics of the modified FCM algorithm are also addressed. Some well known real-world and synthetic datasets are considered for the experiments. In addition to that two remote sensing image datasets are also adopted in this work to illustrate the effectiveness of the proposed FCM over some existing methods. All the obtained results demonstrate that FCM with divergence based proposed similarity measure outperforms three latest FCM algorithms.

1. Introduction

In data mining, the notion of clustering is an essential concept when it comes to unsupervised machine learning algorithm. Clustering is a division of a set of data points into groups by identifying hidden patterns or inherent structures thus the data points in the same cluster have similar properties whereas data points in different groups have highly dissimilar properties. It has wide range of uses for statistical data analysis [1–6] such as performance monitoring for vehicle suspension system [1], mitigating the risk of customer churn [2], fault detection [3], face recognition [4], document clustering [5] and others. Over the past few years, clustering has received considerable amount of attention due to its importance. There is a large amount of clustering algorithms available in literature for finding clusters subject to various constraints. Some of the notable clustering algorithms are k-means [7], DBSCAN [8], Affinity Propagation [9], Normalize Cuts [10] and so on. However, the performance of a clustering algorithm depends on application. Moreover, clustering results rely heavily on similarity measure. In this study, perhaps one of the most widely explored algorithms; fuzzy c-means (FCM) clustering is considered. The purpose of this research is to know the importance of similarity measure in clustering. Thus, clustering criteria is same throughout this work. In general Euclidean similarity measure is frequently used in clustering to minimize the mean squared distance from each data point to its nearest center. But, the Euclidean distance cannot always find more accurate cluster boundaries. In recent years, it is observed that investigators have been replacing the traditional Euclidean distance with the help of non-linear similarity measures to identify more accurate cluster boundaries. Few of them do not follow all the metric properties especially the triangle inequality property [11–14]. General Bregman divergence as a similarity measure was integrated with the k-means to improve the performance of traditional k-means [11]. Interested readers can go through [15–19] to know more about divergence based similarity measures used in clustering. The main contributions of this study include the following:

- A notion of similarity measure is proposed, which is derived from Jeffrey-divergence.
- Various properties of the Jeffrey-divergence based proposed similarity measure are explained.
- The proposed similarity measure is integrated with the traditional FCM clustering algorithm.
- The proof of convergence theorem of the modified FCM algorithm is also discussed.

* Corresponding author at: PDPM Indian Institute of Information Technology, Design and Manufacturing, Jabalpur, 482005, India.
E-mail address: ayan@iiitdmj.ac.in (A. Seal).

https://doi.org/10.1016/j.asoc.2019.106016
1568-4946/

- Simulations are done on nine real and four synthetic datasets to show the performance of the proposed similarity measure over three well-known distance metrics.
- Null hypothesis significance testing is conducted to say the proposed similarity measure outperforms some of the state-of-the-art similarity measures.

The remainder of the work is structured as follows. Section 2 presents a brief overview of FCM. Definition of the proposed similarity measure and its various properties are discussed in Section 3. Section 4 presents the modified FCM algorithm. Simulated results and discussion are illustrated in Section 5. Finally, Section 6 addresses concluding remarks and future scope.

2. Clustering

Formal definition of clustering is described in this section. A brief look of the conventional FCM is also illustrated since comparative study is done between traditional FCM and proposed one.

2.1. Basic principle

Let us consider a set of data points $O[= o_1, o_2, \ldots, o_m]$. Clustering is splitting of O into c groups of similar data points, $K[= (\kappa_1, \kappa_2, \ldots, \kappa_c)]$ thus the degree of association within group is strong whereas bonding is weak between different clusters, where $1 < c < m$. Here, data point, o_i is expressed using d-dimensional feature vector, v_i in R^n_+, where i varies from 1 to m. Mathematically, clustering problem is addressed as follows:

$$\kappa_i \neq \phi \quad for \quad i = 1, \ldots, c,$$
$$\kappa_i \cap \kappa_j = \phi \quad for \quad i = 1, \ldots, c; j = 1, \ldots, c \quad and \quad i \neq j,$$
$$\cup_{i=1}^{c} \kappa_i = K$$

2.2. Fuzzy c-means

In 1973, J. C. Dunn introduced FCM [20] and later on J. C. Bezdek extended it in 1981 [21]. FCM finds groups by minimizing the energy function, $E_f(D, K)$, which is stated in Eq. (1).

$$E_f(K, D; O) = \sum_{y=1}^{m}\sum_{x=1}^{c}(\lambda_{xy})^f \|v_y - \kappa_x\|^2, \quad 1 \leq f < \infty, \tag{1}$$

where f is a real quantity, which represents the fuzziness coefficient. Furthermore, the affect of membership grades in the performance index can be controlled by f. If we increase f then the partition becomes fuzzier. Investigators showed that the FCM converges for any value of f in between 1 and ∞. The λ_{xy} stands for the degree of belongingness/membership of o_y in group K stored in $D(O)_{(c\times m)}$. In case of crisp partitioning, $\lambda_{xy} = 0$ whereas $\lambda_{xy} = 1$ only when o_y belongs to κ_x. The o_y and κ_x are d-dimensional vectors. The former one is the yth data point whereas the later one is the representative of cluster center. The $\|\cdot\|$ is the distance between the center of a cluster and any data points. The energy function depends on K and D, subject to criteria, which are displayed in Eqs. (2) and (3).

$$\sum_{x=1}^{c}\lambda_{xy} = 1, \quad y = 1, 2, \ldots, m, \tag{2}$$

where $\lambda_{xy} \in [0, 1], x = 1, 2, \ldots, c \& y = 1, 2, \ldots, m$.

$$0 < \sum_{y=1}^{m}\lambda_{xy} < c, \quad x = 1, 2, \ldots, c \tag{3}$$

Fuzzy partition undergoes an iterative optimization approach with the update of λ_{xy} and κ_x by Eqs. (4) and (5) respectively.

$$\lambda_{xy}^{(i+1)} = \frac{1}{\sum_{l=1}^{c}\left(\frac{v_y - \kappa_x^{(i)}}{v_y - \kappa_l^{(i)}}\right)^{\frac{2}{f-1}}} \tag{4}$$

$$\kappa_x^{(i+1)} = \frac{\sum_{y=1}^{m}\left[\lambda_{xy}^{(i+1)}\right]^f \cdot v_y}{\sum_{y=1}^{m}\left[\lambda_{xy}^{(i+1)}\right]^f} \tag{5}$$

These updates happen till $max_{xy}\{|\lambda_{xy}^{(i+1)} - \lambda_{xy}^{(i)}|\} < \epsilon$, where ϵ is known as terminating condition. The value of ϵ would be in between 0 and 1. This method tends to meet to a local minimum or a saddle point of $E_f(D, K)$.

3. The proposed similarity measure and its properties

In this section, we present the definition of the proposed similarity measure and its various properties.

Definition 3.1. Eq. (6) is used to estimate Jeffreys-divergence, J_m, which is stated over a set of all positive definite matrices of size $m \times m$ [14].

$$\partial(A, B) = (A - B)(\log(A) - log(B)) \tag{6}$$

where, $|A|$ = determinant of A. An injective function is defined as $\psi : \mathbb{R}^n_+ \to J_m$ thus $\psi(O) = \text{diag}(o_1, o_2, \ldots, o_m)$, where $O = (o_1, o_2, \ldots, o_m) \in \mathbb{R}^n_+$ is a real positive vector. The description of proposed similarity measure includes the following:

Definition 3.2. The similarity between any two data points, $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n_+$, can be expressed as a mapping $S : \mathbb{R}^n_+ \times \mathbb{R}^n_+ \to \mathbb{R}_+ \cup \{0\}$ that can be thought as Eq. (7).

$$S(\mathbf{a}, \mathbf{b}) = \partial(\psi(\mathbf{a}), \psi(\mathbf{b})) \tag{7}$$

Some metric properties of the proposed similarity measure includes the following.

Proposition 3.1. $S(\mathbf{a}, \mathbf{b}) = S(\mathbf{b}, \mathbf{a})$

Proof. $S(\mathbf{a}, \mathbf{b}) = \partial(\psi(\mathbf{a}), \psi(\mathbf{b})) = \partial(\psi(\mathbf{b}), \psi(\mathbf{a})) = S(\mathbf{b}, \mathbf{a})$

Proposition 3.2. $S(\mathbf{a}, \mathbf{b}) \geq 0$ *and* $S(\mathbf{a}, \mathbf{b}) = 0$ *iff* $\mathbf{a} = \mathbf{b}$

Proof. $S(\mathbf{a}, \mathbf{b}) = \partial(\psi(\mathbf{a}), \psi(\mathbf{b})) \geq 0$ and $S(\mathbf{a}, \mathbf{b}) = 0$ iff $\partial(\psi(\mathbf{a}), \psi(\mathbf{b})) = 0$ iff $\psi(\mathbf{a}) = \psi(\mathbf{b})$ iff $\mathbf{a} = \mathbf{b}$

Thus, S is a similarity measure on $\mathbb{R}^n_+$, but it is not a distance metric because it does not obey the triangle inequality property. So, the proposed similarity measure can be represented as $S(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{m}\partial(a_i, b_i)$. At this time, we investigate some of the properties of the proposed similarity measure.

Theorem 3.1. *The proposed similarity measure is not a Bregman divergence.*

Proof. Theorem 3.1 can be proved by assuming the opposite. Supposing the proposition that is *the proposed similarity measure was a Bregman divergence* $S(\mathbf{a}, \mathbf{b})$ is false if $S(\mathbf{a}, \mathbf{b})$ is convex in $\mathbf{a}$. The S can also be articulated by Eq. (8).

$$S(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{m}(a_i - b_i)(\log(a_i) - \log(b_i)) \tag{8}$$

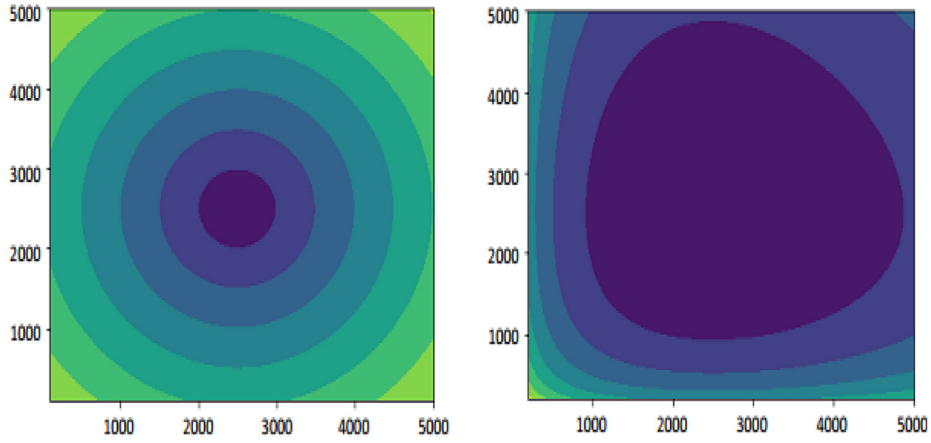

Fig. 1. Contour plot of norm ball for Euclidean distance and the proposed similarity measure.

We can differentiate both sides of Eq. (8) w.r.t a_i to obtain the following expression:

$$\frac{\partial S}{da_i} = 1 - \frac{b_i}{a_i} + \log(a_i) - \log(b_i)$$

$$\frac{\partial^2 S}{da_i da_j} = 0 \text{ when } i \neq j \text{ otherwise,}$$

$$\frac{\partial^2 S}{\partial a_i^2} = \frac{b_i}{a_i^2} + \frac{1}{a_i}$$

In the range of $\{-\infty, -1\} \cup \{0, 1\}$, $S(\mathbf{a}, \mathbf{b})$, the values of $\frac{\partial^2 S}{\partial a_i^2} < 0$. Hence, $S(\mathbf{a}, \mathbf{b})$ is not convex in $\mathbf{a}$. In other words, the proposed similarity measure is not a Bregman divergence.

Theorem 3.2. *$S(\mathbf{e} \circ \mathbf{a}, \mathbf{e} \circ \mathbf{b}) = \mathbf{e}S(\mathbf{a}, \mathbf{b})$ for $\mathbf{e} \in \mathbb{R}_+^n$, where $\mathbf{e} \circ \mathbf{a}$ designates as the Hadamard product between $\mathbf{e}$ and $\mathbf{a}$.*

Proof. We know, $(\mathbf{e} \circ \mathbf{a}) = (e_1 a_1, e_2 a_2, \ldots, e_m a_m)$. So, $\delta(e_i a_i, e_i a_i) = (e_i a_i - e_i b_i)(\log(e_i a_i) - \log(e_i b_i)) = e_i(a_i - b_i)(\log e_i + \log a_i - \log e_i - \log b_i) = e_i(a_i - b_i)(\log a_i - \log b_i)$ $\sum_{i=1}^m \delta(e_i a_i, e_i b_i) = \sum_{i=1}^m e_i \delta(a_i, b_i)$ implying $S(\mathbf{e} \circ \mathbf{a}, \mathbf{e} \circ \mathbf{b}) = \mathbf{e}S(\mathbf{a}, \mathbf{b})$

Theorem 3.3. *The proposed similarity measure is f-divergence.*

Proof. A divergence is known as f-divergence if it can be stated as $\phi(t) = \mathbf{a}\phi(\frac{\mathbf{b}}{\mathbf{a}})$, where $t = \frac{\mathbf{b}}{\mathbf{a}}$

The similarity between $\mathbf{a} \in \mathbb{R}_+^n$ and $\mathbf{b} \in \mathbb{R}_+^n$ be given as $S(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^m (a_i - b_i)(\log(a_i) - \log(b_i))$ putting $t_i = \frac{b_i}{a_i}$

$$S(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^m (a_i - a_i t_i)(\log(a_i) - \log(a_i t_i))$$

$$= \sum_{i=1}^m a_i(1 - t_i)(\log(a_i) - \log(a_i) - \log(t_i))$$

$$= \sum_{i=1}^m a_i(1 - t_i)(-\log(t_i))$$

$$= \sum_{i=1}^m a_i(1 - t_i)(\log(\frac{1}{t_i}))$$

$$\sum_{i=1}^m \phi(t) = \sum_{i=1}^m a_i \phi(\frac{b_i}{a_i})$$

Since, $S(\mathbf{a}, \mathbf{b})$ can be expressed as $\sum_{i=1}^m a_i \phi(\frac{b_i}{a_i})$. Thus, the proposed similarity measure is f-divergence.

Remark 3.1. Let us see another imperative characteristic of the proposed similarity measure. Fig. 1 shows the profile of the norm-balls in $\mathbb{R}^2$ surrounding the point (5000,5000) for Euclidean distance (Fig. 1a) and the proposed similarity measure (Fig. 1b). It is observed from Fig. 1 that the norm-ball of Euclidean distance is like concentric circle whereas the proposed similarity measure is somewhat like distorted ovals. It is also clear from Fig. 1b that contour lines resemble each other as we move towards the origin i.e. (0,0). Thus, we arrive at a conclusion that the Jeffreys divergence between two points are higher when they are near to origin and it decreases while they are away from the origin. On the other hand, Euclidean distance between two points are same irrespective of the position. For example, the Euclidean distance and J-divergence between (3,3) and (5,5) is 2.82 and 2.043 respectively and for points (1003,1003) and (1005,1005) they are 2.82 and 0.0079 respectively. Sometimes, this property would be useful when clusters having different densities and sizes.

4. Modified fuzzy c-means with the proposed similarity measure

The FCM with the proposed similarity measure achieves grouping by solving Eq. (9).

$$\min_{\substack{K=(\kappa_1,\kappa_2,\ldots,\kappa_C)\in R^{m\times c} \\ D\in M}} E_f(K, D; O) = \sum_{y=1}^{m}\sum_{x=1}^{c}(\lambda_{xy})^f S(o_y, \kappa_x), \quad 1 \leq s < \infty \tag{9}$$

where,

$$M = \left\{ D = [\lambda_{xy}]_{\substack{x=1,2,\ldots,c \\ y=1,2,\ldots,m}} \middle| \lambda_{xy} \in [0, 1], \sum_{x=1}^{c} \lambda_{xy} = 1, \sum_{y=1}^{m} \lambda_{xy} > 0 \right\} \tag{10}$$

Exact solution of Eq. (10) does not exist [22]. An alternating optimization method exists in literature to find a solution, which is as follows:

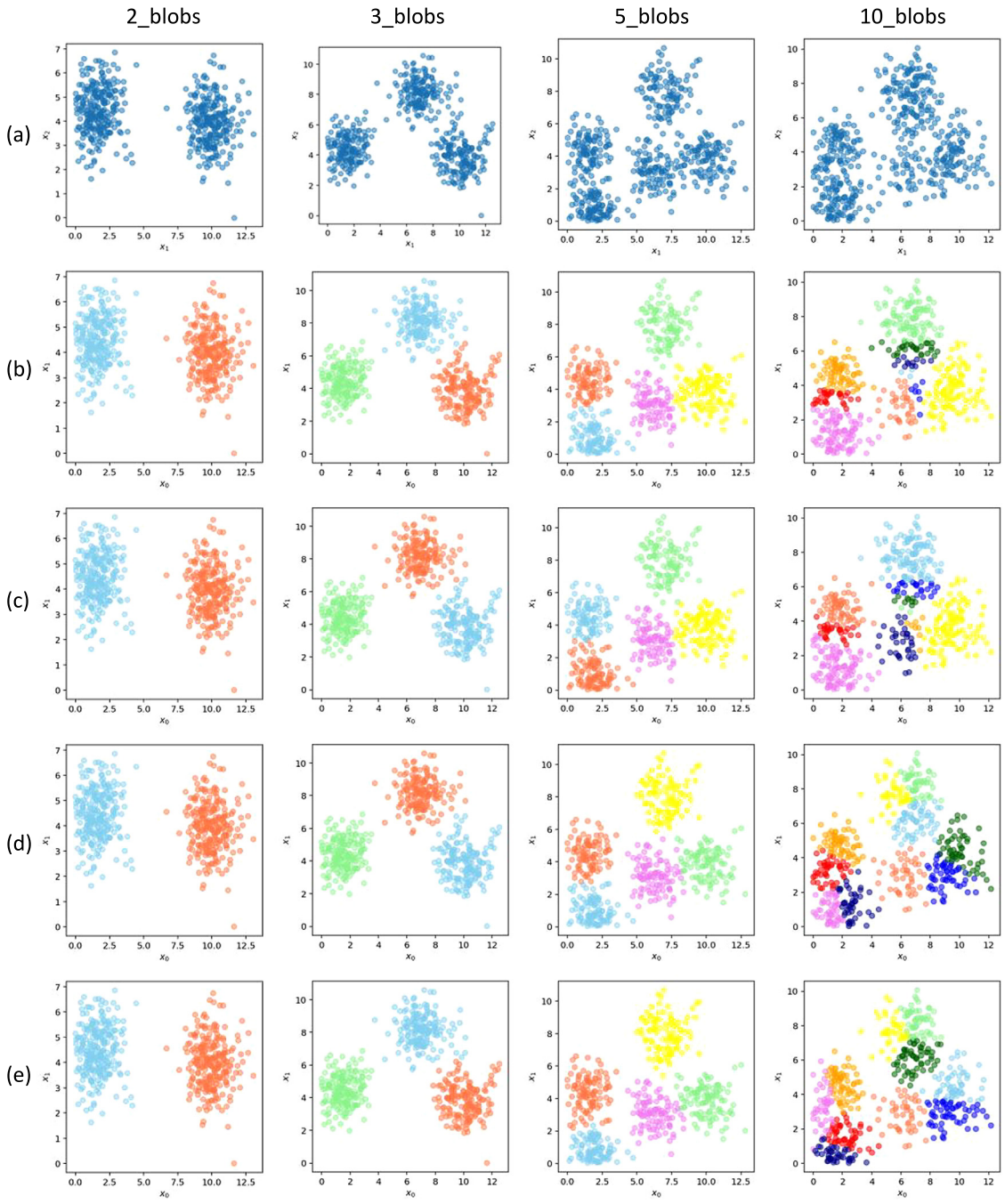

Fig. 2. Clustering results: (a) original data distribution (b) groups using M_1 (c) groups using M_2 (d) groups using M_3 and (e) groups using M_4.

Theorem 4.1. *Supposing* $\tau_y = \left\{x | x \in [1, c], o_y = \kappa_x^{(i)}\right\}$ *where, i is the epoch number. Eq. (11) is the alternative form of Eq. (4) whereas Eq. (5) is alike. Both the Eqs. are necessary in alternating optimization algorithm for the proof of convergence of* E_f *[23].*

$$\lambda_{xy}^{(i+1)} = \begin{cases} \left(\sum_{l=1}^{c}\left[\dfrac{S(o_y, \kappa_x^{(i)})}{S(o_y, \kappa_l^{(i)})}\right]^{\frac{2}{f-1}}\right)^{-1}, & \text{if } \tau_y = 0 \\ \dfrac{1}{|\tau_y|}, & \text{if } \tau_y \neq 0 \text{ and } x \in \tau_y \\ 0, & \text{if } \tau_y \neq 0 \text{ and } x \notin \tau_y \end{cases} \tag{11}$$

The FCM criterion in Eq. (9) can be stated through reduced unconstrained FCM criterion in Theorem 4.2.

Theorem 4.2. *The reduced FCM criterion appears in Eq. (12), which is alike [22,24,25]*

$$\min_{K \in R^{m \times c}} E_f'(K; O) = \sum_{y=1}^{m}\left[\sum_{x=1}^{c} S(o_y, \kappa_x)^{\frac{2}{1-f}}\right]^{1-f} \tag{12}$$

K^* is a saddle point of E_f' when (K^*, D^*) is a saddle point of E_f. $(K^*, F(K^*))$ is a saddle point of E_f when K^* is a saddle point of E_f', where $F : R^{m \times c} \rightarrow M$ and $F(K) = D$ with each λ_{xy} calculated by Eq. (11). The proof of theorem is beyond the scope of this study. Interested readers are refereed to know in detail [24,25].

Convergence FCM: Supposing $f(\gamma_1, \gamma_2, \ldots, \gamma_k) = (\sum_{x=1}^{c} \gamma_x^s)^{1/s}$, where $s = \frac{1}{1-f} < 0$. Then Eq. (12) can be expressed as

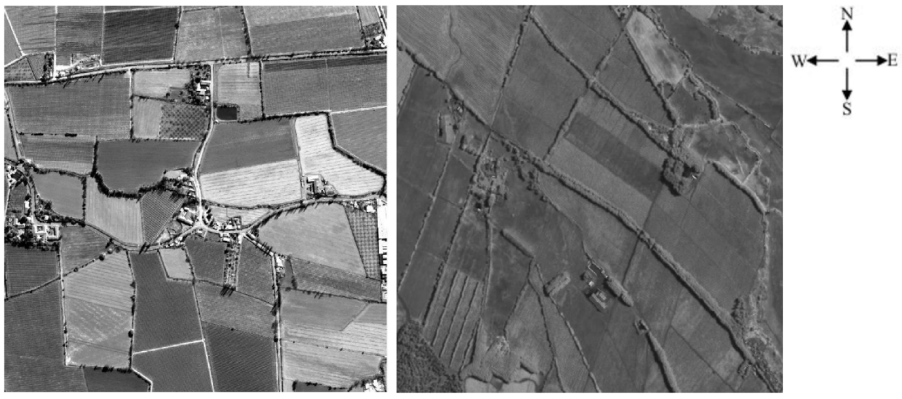

Fig. 3. From left to right: Panchromatic image of DS_{12} and DS_{13}.

$$E'_f(\kappa_1, \kappa_2, \ldots, \kappa_c; O) = \sum_{y=1}^{m} f(\gamma_{\kappa 1}, \gamma_{\kappa 2}, \ldots, \gamma_{\kappa x})|_{\gamma_{xy}=S(o_y,\kappa_x)^2} \tag{13}$$

Lemma 4.1 can be defined through Eq. (13), where the RHS of Eq. (14) is called as a majorant of E'_f.

Lemma 4.1 (*Majorant of E'_f*)**.**

$$E'_f(\kappa_1, \kappa_2, \ldots, \kappa_c; O) \leq maj^e E'_f = E'_f(\kappa_1^{(i)}, \kappa_2^{(i)}, \ldots, \kappa_c^{(i)}; O) + \sum_{y=1}^{m}\sum_{x=1}^{c} \frac{df}{d\gamma_{xy}}|_{(i)}\left(S(o_y, \kappa_x)^2 - S(o_y, \kappa_x^i)^2\right) \tag{14}$$

where, the derivative is taken at $\kappa_1^{(i)}, \kappa_2^{(i)}, \ldots, \kappa_c^{(i)}$.

Proof. L. Gröll el at. established the fact that $f(\gamma_1, \gamma_2, \ldots, \gamma_c)$ is concave [22]. Hence,

$$f(\gamma_1, \gamma_2, \ldots, \gamma_c) \leq f(\rho_1, \rho_2, \ldots, \rho_c) + \sum_{x=1}^{c} \frac{df}{d\gamma_{x_{\rho x}}}(\gamma_x - \rho_x) \tag{15}$$

Eq. (16) can be obtained by replacing γ_x and ρ_x using γ_{xy} and ρ_{xy} respectively and considering the sum over all y.

$$\sum_{y=1}^{m} f(\gamma_{1y}, \gamma_{2y}, \ldots, \gamma_{cy}) \leq \sum_{y=1}^{m} f(\rho_{1y}, \rho_{2y}, \ldots, \rho_{cy}) + \sum_{y=1}^{m}\sum_{x=1}^{c} \frac{df}{d\gamma_{xy_{\rho xy}}}(\gamma_{xy} - \rho_{xy}) \tag{16}$$

Eq. (17) can be inherited by assigning the value of $\gamma_{xy} = d(o_y, \kappa_x)^2$ and $\rho_{xy} = d(o_y, \kappa_x^{(i)})^2$ as well as Eq. (14).

$$E'_f(\kappa_1, \kappa_2, \ldots, \kappa_k; O) \leq E'_f(\kappa_1^{(i)}, \kappa_2^{(i)}, \ldots, \kappa_c^{(i)}; O) + \sum_{y=1}^{m}\sum_{x=1}^{c} \frac{df}{d\gamma_{xy}}|_{(i)}\left(S(o_y, \kappa_x)^2 - S(o_y, \kappa_x^i)^2\right) \tag{17}$$

Each majorant got from Eq. (14) is a global majorant. In other words, a majorant is global along a random search direction. The minimizers of global and directional majorants are identical iff the search direction passes the global minimizer of the global majorant.

Theorem 4.3 (*Steepest Descent Algorithm for an Alternating Optimization*)**.** *If the step length is adjusted by the majorization principle, Eq. (14), then the sequences $\kappa_x^{(i+1)}$ appeared in the alternating optimization algorithm in the form of Eqs. (11) and (5) and the sequences of a steepest descent algorithm applied to Eq. (12) are identical.*

Proof. All the coefficients, $\frac{df}{d\gamma_{xy}}$, of the rigidly convex terms $d(o_y, \kappa_x)^2$ are non-negative.

$$\begin{aligned}\frac{df}{d\gamma_{xy}} &= \frac{d}{d\gamma_{xy}}\left[\sum_{x=1}^{c}\gamma_{xy}^{s}\right]^{\frac{1}{s}} = \left[\sum_{x=1}^{c}\gamma_{xy}^{s}\right]^{\frac{1}{s-1}}\gamma_{xy}^{s-1} \\ &= \left[\sum_{l=1}^{c}\left(\frac{1}{\gamma_{ly}}\right)^{-s}\right]^{\frac{1}{s-1}}(\gamma_{xy}^{-s})^{\frac{1}{s-1}} = \left[\sum_{l=1}^{c}\frac{\gamma_{xy}^{-s}}{\gamma_{ly}^{-s}}\right]^{\frac{1}{s-1}} \\ &= \left[\left(\sum_{l=1}^{c}\left[\frac{S(o_y,\kappa_x)^2}{S(o_y,\kappa_l)^2}\right]^{\frac{1}{f-1}}\right)^{-1}\right]^{f} = (\lambda_{xy})^f \geq 0\end{aligned} \tag{18}$$

So, the majorant is convex w.r.t. each κ_x. Furthermore, majorant is convex since one or more coefficients corresponding to each κ_x is non-negative. Thus, the one and only minimizer, κ_x^o, of the majorant by the first-order both enough and sufficient condition is

$$\begin{aligned}\nabla_{\kappa_x} maj^{(i)} &= E'_f(\kappa_1, \kappa_2, \ldots, \kappa_c; O)|_{\kappa_x=\kappa_x^o} \\ &= -2\sum_{y=1}^{m}\frac{df}{d\gamma_{xy}}|_{(i)}(o_y - \kappa_x^o) = 0, a = 1, 2, \ldots, c\end{aligned} \tag{19}$$

So, the value of $\kappa_x^{(i+1)}$ is

$$\kappa_x^{(i+1)} = \kappa_x^o = \frac{\sum_{y=1}^{m}\frac{df}{d\gamma_{xy}}|_{(i)}o_y}{\sum_{y=1}^{m}\frac{df}{d\gamma_{xy}}|_{(i)}} \tag{20}$$

Eq. (20) would be same as Eq. (5) after replacing $\frac{df}{d\gamma_{xy}}|_{(i)}$ by $(\lambda_{xy}^{(i+1)})^f$. The steepest descent may be calculated by Eq. (21).

$$\kappa_x^{(i+1)} = \kappa_x^{(i)} - \underbrace{\frac{1}{2\sum_{y=1}^{m}\frac{df}{d\gamma_{xy}}|_{(i)}}}_{\text{steplength } \alpha_x^{(i)}} \cdot \underbrace{\left(-2\sum_{b=1}^{m}\frac{df}{d\gamma_{xy}}|_{(i)}(o_y - \kappa_x^{(i)})\right)}_{\nabla_{\kappa_x}E'_f(\kappa_1,\kappa_2,\ldots,\kappa_c;O)|_{\kappa_l=\kappa_l^{(i)},\quad l=1,2,\ldots,c}} \tag{21}$$

Finally, the global minimizer majorants are directions along the direction with steepest descent.

$$\begin{aligned}&\nabla_{\kappa_x} maj^{(i)}E'_f(\kappa_1, \kappa_2, \ldots, \kappa_c; O)|_{\kappa_l=\kappa_l^{(i)}} \\ &= \underbrace{\nabla_{\kappa_x}E'_f(\kappa_1, \kappa_2, \ldots, \kappa_c; O)|_{\kappa_l=\kappa_l^{(i)},\quad l=1,2,\ldots,c}}_{\nabla_{\kappa_x}E_f^{'(i)}}\end{aligned} \tag{22}$$

Table 1
The values of cluster validity indexes namely, ARI, NMI, SI, DI and DBI for synthetic and real-world datasets.

	Dataset	M_1	M_2	M_3	M_4
ARI	DS_1	**1.0**	**1.0**	**1.0**	**1.0**
	DS_2	0.988018	0.982036	0.982036	**0.990105**
	DS_3	0.875980	0.875845	0.898623	**0.927475**
	DS_4	0.463736	0.448327	0.432681	**0.515745**
	DS_5	0.921409	0.902619	0.979932	**1.0**
	DS_6	0.446451	0.431981	0.5476595	**0.597594**
	DS_7	0.577500	0.78239	0.484217	**0.881025**
	DS_8	0.550272	0.550272	0.532513	**0.915032**
	DS_9	0.027349	0.0317099	0.020098	**0.032356**
	DS_{10}	0.528617	0.5391607	0.491424	**0.903304**
	DS_{11}	0.107842	0.107842	0.106418	**0.950622**
	DS_{12}	0.406837	0.3079934	0.407870	**0.648688**
	DS_{13}	0.320471	0.3204717	0.378029	**0.390829**
NMI	DS_1	**1.0**	**1.0**	**1.0**	**1.0**
	DS_2	0.977740	0.966614	0.982036	**0.990181**
	DS_3	0.866303	0.868204	0.898623	**0.906569**
	DS_4	0.654322	0.643144	0.432681	**0.686579**
	DS_5	0.897779	0.879632	0.979932	**0.983390**
	DS_6	0.616544	0.593026	0.5476595	**0.670032**
	DS_7	0.396832	0.53360	0.484217	**0.856987**
	DS_8	0.451882	0.4510854	0.532513	**0.558972**
	DS_9	0.027349	0.020370	0.020098	**0.870295**
	DS_{10}	0.486361	0.5040837	0.491424	**0.841138**
	DS_{11}	0.086286	0.086286	0.106418	**0.915113**
	DS_{12}	0.328378	0.2955786	0.512618	**0.560183**
	DS_{13}	0.019634	0.4407031	0.418665	**0.471018**
SI	DS_1	0.794916	0.794873	0.798578	**0.813145**
	DS_2	0.6931593	0.693228	0.693159	**0.695152**
	DS_3	0.572960	0.571924	0.571924	**0.589525**
	DS_4	0.256123	0.248279	0.248279	**0.654126**
	DS_5	0.577102	0.574844	0.577102	**0.589541**
	DS_6	0.315921	0.246860	0.315921	**0.317859**
	DS_7	0.399785	0.394392	0.399785	**0.399891**
	DS_8	0.410091	0.407459	0.410091	**0.412264**
	DS_9	0.027349	0.369059	0.027349	**0.373494**
	DS_{10}	0.691067	0.690935	0.691067	**0.691234**
	DS_{11}	0.530125	0.530125	0.530125	**0.545125**
	DS_{12}	0.314455	0.390665	0.297673	**0.465552**
	DS_{13}	0.021033	0.025401	0.021698	**0.051497**
DI	DS_1	1.904015	1.914016	1.904015	**1.996216**
	DS_2	1.667249	1.704798	1.667249	**1.742156**
	DS_3	1.25921	1.307801	1.307801	**1.321452**
	DS_4	0.463736	0.220063	0.463736	**0.489652**
	DS_5	1.897049	1.840235	1.840235	**1.989932**
	DS_6	0.444175	0.065613	0.446595	**0.465213**
	DS_7	1.423736	1.364377	1.484217	**1.659659**
	DS_8	1.00112	1.001708	1.532513	**1.567329**
	DS_9	1.257200	**1.297200**	1.020098	1.029349
	DS_{10}	1.300558	1.300383	1.491424	**1.578617**
	DS_{11}	1.397413	1.397412	1.106418	**1.421578**
	DS_{12}	0.392247	0.474642	0.509411	**0.517705**
	DS_{13}	0.391262	0.391262	0.377334	**0.438756**
DBI	DS_1	**0.144604**	0.145413	0.144634	**0.144604**
	DS_2	0.165546	0.156589	0.156546	**0.156546**
	DS_3	0.122504	0.121585	0.121585	**0.120504**
	DS_4	0.187655	0.196655	**0.141915**	**0.141915**
	DS_5	0.172985	0.167492	**0.167452**	**0.167452**
	DS_6	0.507572	**0.253338**	0.513267	**0.253338**
	DS_7	0.496210	0.472901	0.472901	**0.468776**
	DS_8	**0.518281**	0.521241	0.523546	**0.518281**
	DS_9	0.499654	0.499391	0.499391	**0.027349**
	DS_{10}	0.258860	0.257766	0.257766	**0.257680**
	DS_{11}	0.335503	**0.319149**	**0.319149**	**0.319149**
	DS_{12}	0.272598	0.539363	0.448134	**0.136888**
	DS_{13}	0.228156	0.228156	0.210953	**0.202353**

Table 2
The computed p-Values based on clustering validity indexes for Wilcoxon's signed rank test to examine M_4 over M_i, where $1 \leq i \leq 3$.

	Dataset	M_1	M_2	M_3
ARI	DS_1	1	1	1
	DS_2	0.0025	0.0025	0.0025
	DS_3	0.0020	0.002	0.0020
	DS_4	0.0020	0.0020	0.0020
	DS_5	0.0020	0.0020	0.0020
	DS_6	0.0371	0.0440	0.0195
	DS_7	0.0020	0.0020	0.0020
	DS_8	0.0022	0.0020	0.0015
	DS_9	0.0020	0.0020	0.0015
	DS_{10}	0.002	0.002	0.002
	DS_{11}	0.002	0.002	0.002
	DS_{12}	0.0020	0.0015	0.0025
	DS_{13}	0.0025	0.0020	0.0023
NMI	DS_1	1	1	1
	DS_2	0.0025	0.0025	0.0025
	DS_3	0.0020	0.002	0.0020
	DS_4	0.0020	0.0020	0.0020
	DS_5	0.0020	0.0020	0.0020
	DS_6	0.0025	0.0025	0.0025
	DS_7	0.0020	0.002	0.0020
	DS_8	0.0025	0.0025	0.0025
	DS_9	0.0025	0.0025	0.0025
	DS_{10}	0.0020	0.002	0.0020
	DS_{11}	0.0025	0.0025	0.0025
	DS_{12}	0.0020	0.0020	0.0025
	DS_{13}	0.0015	0.0020	0.0035
SI	DS_1	0.0020	0.0020	0.0020
	DS_2	0.0020	0.0020	0.0020
	DS_3	0.0020	0.0020	0.0020
	DS_4	0.0025	0.0020	0.0020
	DS_5	0.0020	0.0020	0.0020
	DS_6	0.0371	0.0840	0.0195
	DS_7	0.0022	0.0020	0.0020
	DS_8	0.0020	0.0025	0.0020
	DS_9	0.0020	0.0015	0.0020
	DS_{10}	0.0020	0.0020	0.0020
	DS_{11}	0.0020	0.0020	0.0020
	DS_{12}	0.0015	0.0010	0.0005
	DS_{13}	0.0020	0.0020	0.0020
DI	DS_1	0.0020	0.0020	0.0020
	DS_2	0.0020	0.0020	0.0020
	DS_3	0.0020	0.0020	0.0020
	DS_4	0.0025	0.0020	0.0020
	DS_5	0.0020	0.0020	0.0020
	DS_6	0.0025	0.0020	0.0020
	DS_7	0.0022	0.0020	0.0020
	DS_8	0.0020	0.0025	0.0020
	DS_9	0.0020	0.0015	0.0020
	DS_{10}	0.0025	0.0020	0.0020
	DS_{11}	0.0022	0.0022	0.0020
	DS_{12}	0.0020	0.0002	0.0005
	DS_{13}	0.0015	0.0020	0.0005
DBI	DS_1	0.0020	0.0020	0.0020
	DS_2	0.0020	0.0020	0.0020
	DS_3	0.0015	0.0020	0.0022
	DS_4	0.0020	0.0020	0.0020
	DS_5	0.0020	0.0020	0.0020
	DS_6	0.0025	0.0020	0.0020
	DS_7	0.0022	0.0020	0.0020
	DS_8	0.0020	0.0025	0.0020
	DS_9	0.0020	0.0015	0.0020
	DS_{10}	0.0020	0.0025	0.0020
	DS_{11}	0.0020	0.0020	0.0020
	DS_{12}	0.0150	0.0005	0.0002
	DS_{13}	0.0020	0.0015	0.0020

At this time, the convergence properties can be easily established by aforementioned optimization theory.

Corollary 4.1 (*Global Convergence of Reduced FCM*)**.** *The reduced FCM states in Eq. (9) that it converges to a saddle point globally.*

Proof. Using Lemma 4.1 $E_f^{'(i)} - E_f^{'(i+1)} \geq E_f^{'(i)} - maj^{(i)}E_f'\left(\kappa_1^{(i+1)}, \kappa_2^{(i+1)}, \ldots, \kappa_c^{(i+1)}; O\right) = \sum_{y=1}^{m}\sum_{x=1}^{c}\frac{df}{d\gamma_{xy}}|_{(i)}\left(S\left(o_y, \kappa_x^{(i)}\right)^2 - S\left(o_y, \kappa_x^{(i)} + \alpha_x^{(i)} \cdot \nabla_{\kappa_x} E_f^{'(i)}\right)^2\right) = \sum_{y=1}^{m}\sum_{x=1}^{c}\frac{df}{d\gamma_{xy}}|_{(i)}\left(-2\alpha_x^{(i)}\left(o_y - \kappa_x^{(i)}\right)^T \nabla_{\kappa_x} E_f^{'(i)} -$

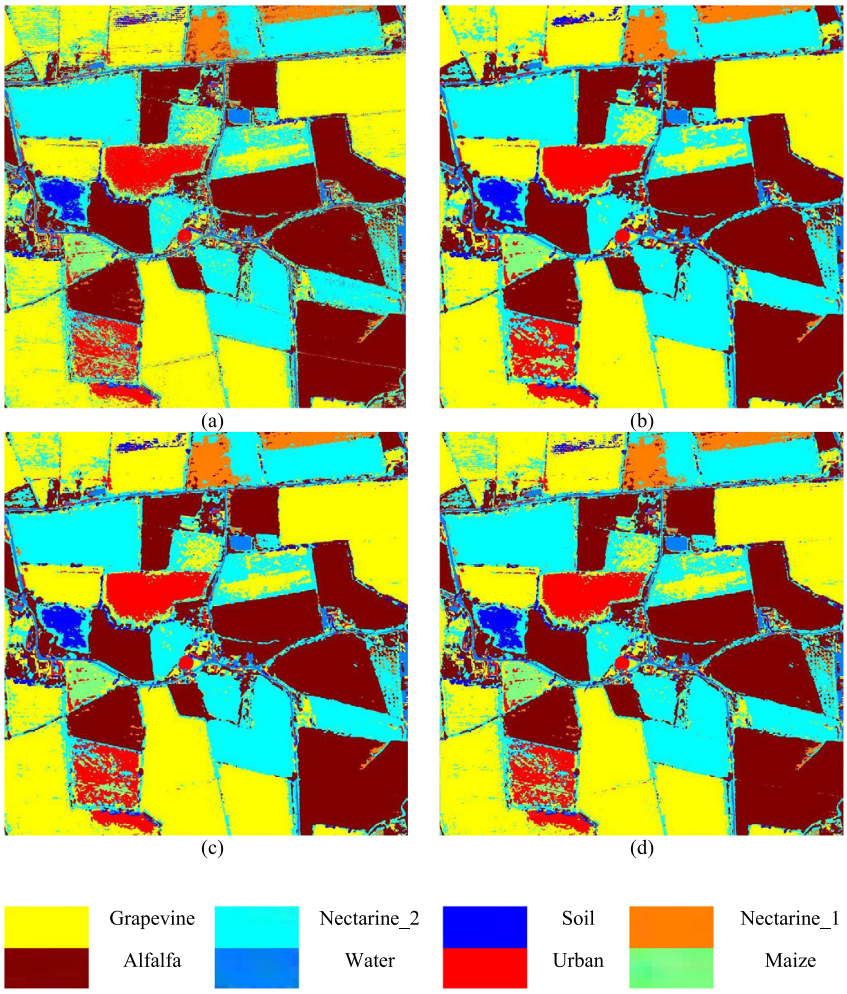

Fig. 4. Clustering results: (a) the output of DS_{12} using M_1 (b) the output of DS_{12} using M_2 (c) the output of DS_{12} using M_3 (d) the output of DS_{12} using M_4.

$\left(\alpha_x^{(i)}\right)^2 \left|\left| \nabla_{\kappa_x} E_f^{'(i)} \right|\right|_2^2\Big) = \sum_{x=1}^{c}\Big(-2\alpha_x^{(i)} \sum_{y=1}^{m} \frac{df}{d\gamma_{xy}}|_{(i)} \left(o_y - \kappa_x^{(i)}\right)^T \nabla_{\kappa_x} E_f^{'(i)} - \sum_{y=1}^{m} \frac{df}{d\gamma_{xy}}|_{(i)} \left(\alpha_x^{(i)}\right)^2 \left|\left| \nabla_{\kappa_x} E_f^{'(i)} \right|\right|_2^2\Big) = \sum_{x=1}^{c}\Big(\alpha_x^{(i)} \left|\left| \nabla_{\kappa_x} E_f^{'(i)} \right|\right|_2^2 - 2\frac{1}{\alpha_x^{(i)}} \left(\alpha_x^{(i)}\right)^2 \left|\left| \nabla_{\kappa_x} E_f^{'(i)} \right|\right|_2^2\Big)$ where, $\nabla_{\kappa_x} E_f^{'(i)} = -2 \sum_{y=1}^{m} \frac{df}{d\gamma_{xy}}|_{(i)} \left(o_y - \kappa_x^{(i)}\right)$ and $\alpha_x^{(i)} = \frac{1}{2\sum_{y=1}^{m} \frac{df}{d\gamma_{xy}}|_{(i)}} = \sum_{x=1}^{c} \frac{\alpha_x^{(i)}}{2} \left|\left| \nabla_{\kappa_x} E_f^{'(i)} \right|\right|_2^2 \geq \sum_{x=1}^{c} \frac{1}{4m} \left|\left| \nabla_{\kappa_x} E_f^{'(i)} \right|\right|_2^2$ since $\sum_{y=1}^{m} \frac{df}{d\gamma_{xy}}|_{(i)} \leq m$ and hence $\alpha_x^{(i)} \geq \frac{1}{2m}$

So, $E_f^{'(i)} \geq E_f^{'(i+1)}$. The boundedness of E_f' follows $\lim_{i\to\infty}\left(E_f^{'(i)} - E_f^{'(i+1)}\right) = 0$. The RHS of the inequality is 0 if the LHS approaches to 0 because it is bounded by 0. Convergence to a saddle point can be achievable from $\lim_{i\to\infty} \left|\left| \nabla_{\kappa_x} E_f^{'(i)} \right|\right|_2^2 = 0$ $\forall a$, and so, $\lim_{i\to\infty} \nabla_{\kappa_x} E_f^{'(i)} = 0$. Only local minimizers or saddle points appear as limit points because $\{E_f^{'(i)}\}$ is a non-decreasing sequence.

Corollary 4.2 (*Local Convergence of Reduced FCM*)**.** *There have neighborhoods $Z(\kappa_1^*, \kappa_2^*, \ldots, \kappa_c^*)$ if $K = (\kappa_1^*, \kappa_2^*, \ldots, \kappa_c^*)$ is a rigid local minimizer of E_f' so as if a beginning point $K^{(0)} = (\kappa_1^{(0)}, \kappa_2^{(0)}, \ldots, \kappa_c^{(0)})$ is selected from the neighborhood, the FCM algorithm converges to $K^* = (\kappa_1^*, \kappa_2^*, \ldots, \kappa_c^*)$.*

Proof. Since the FCM algorithm is a globally convergent gradient approach using Corollary 4.1., the popular Capture Theorem can be employed [26].

Corollary 4.3 (*Convergence Rate of FCM*)**.** *FCM coincides linearly near a non-singular local minimum with a positive definite Hessian matrix.*

Proof. The proof of Corollary 4.3. employs a Taylor expansion of E_f' and the popular convergence rate theorem for quadratic functions [26].

5. Experimental results and discussion

5.1. Description of datasets

Nine real-world and four synthetic datasets are adopted to conduct experiments, where synthetic datasets contain 2_blobs (DS_1), 3_blobs (DS_2), 5_blobs (DS_3), and 10_blobs (DS_4). Here, the number of blobs means the number of clusters, which are synthesized by normal distributions. The first row of Fig. 2 depicts the data points distributions of DS_1, DS_2, DS_3 and DS_4. Seven real-world datasets namely, Iris, Glass, Cleveland, Mammography, Breast Cancer Wisconsin, Appendicitis and Bank Note Authentication are mustered from the UCI Machine Learning Repository [27]

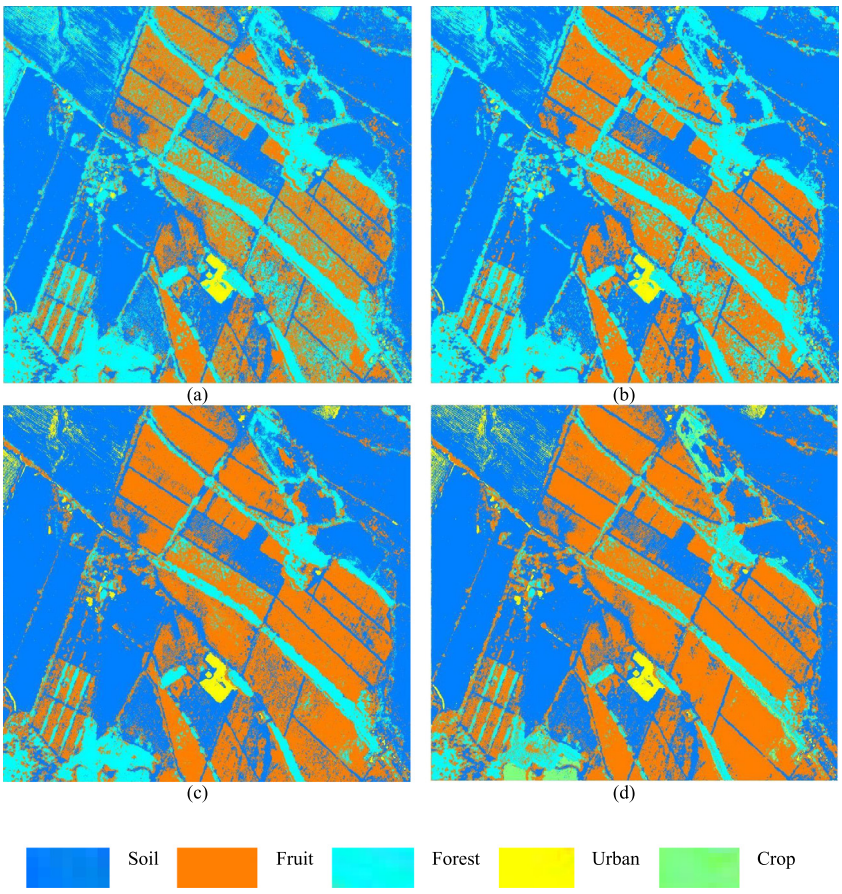

Fig. 5. Clustering results: (a) the output of DS_{13} using M_1 (b) the output of DS_{13} using M_2 (c) the output of DS_{13} using M_3 (d) the output of DS_{13} using M_4.

and Keel Repository [28], which are designated as DS_5, DS_6, DS_7, DS_8, DS_9, DS_{10} and DS_{11} respectively.

Two sets of panchromatic and multi-spectral satellite images are also adopted, which are labeled as DS_{12} and DS_{13}. These are acquired by worldview-2 sensor with a scale 1:25,000. Interested readers are referred to [29] to know more about these datasets. Panchromatic images of DS_{12} and DS_{13} are shown in Fig. 3. The DS_{12} and DS_{13} have 8 and 5 different land covers.

5.2. Clustering validity measurement indexes

In clustering, measuring the "goodness" of the output clusters is a fundamental problem. Validity indexes help to measure the concept of goodness. Mathematical definition of a validity index is as follows: Supposing O has m data points. The O can be partitioned into c-groups viz., $O_1, O_2, \ldots, O_c$ using a suitable clustering algorithm. The V_1, V_2, ..., V_c are the values of the validity indexes of $O_1, O_2, \ldots, O_c$ respectively. The $V_{h1} \geq V_{h2} \geq \cdots \geq V_{hc}$ will depict $O_{h1} \uparrow O_{h2} \uparrow \ldots \uparrow O_{hc}$, for any permutation $h1, h2, \ldots, hc$ of $\{1, 2, \ldots, c\}$, where $O_i \uparrow O_j$ denotes cluster O_i is a better than O_j in some perception [30]. In general, cluster validity indexes are broadly classified into two classes, viz., external and internal. The former one matches predicted cluster labels produced by a clustering algorithm with the actual class labels that are supplied externally. Two external validity indexes, viz., Adjusted Rand Index (ARI) [31] and Normalized Mutual Information (NMI) [32] are considered in this work. No external information is required for internal validity index to measure the "goodness" of clusters. Dunn index (DI) [30], Davies Boulden Index (DBI) [30], and Silhouette index (SI) [33] are adopted as internal cluster validity indexes to quantify the cohesiveness of the obtained clusters. The range of NMI and ARI varies from 0 to 1. One indicates that two groups of data points are alike whereas 1 says that both the groups are completely dissimilar. Matching is done between the predicted partition by a clustering algorithm and the ground truth. On the other hand, internal clustering evaluation metrics estimate the closeness of a data point to its own group (cohesion) compared to other groups (separation). The range of SI varies from -1 to $+1$, where a value close to 1 denotes that the data points are similar in its own group and poorly dissimilar to neighbor groups. A higher DI and lower DBI represent better clustering.

5.3. Computational protocols

Four simulations are conducted on all datasets using four different similarity measures, which are as follows:

- M_1: FCM with Euclidean distance
- M_2: Minkowski weighted FCM
- M_3: Weighting in FCM
- M_4: FCM with the proposed similarity measure

Performance comparison: It confirms that identical randomly selected centroids are considered for each of the method in order to estimate the values of ARI, NMI, SI, DI, and DBI to maintain the consistency in results. The performance of each method does not depend on the selection of initial set of centroids. However, the performance depends on the method. The same method is executed 10 times on each dataset to list out clustering performances

Table 3
The computed p-Values based on clustering validity indexes for Wilcoxon's ranksum test to examine M_4 over M_i, where $1 \leq i \leq 3$.

	Dataset	M_1	M_2	M_3
ARI	DS_1	1	1	1
	DS_2	0.0079	0.0079	0.0079
	DS_3	0.0020	0.0020	0.0020
	DS_4	1.5938e−05	1.5938e−05	1.5938e−05
	DS_5	1.1758e−05	1.5938e−05	1.3659e−05
	DS_6	1.5938e−05	1.1758e−05	1.5938e−05
	DS_7	1.4523–05	1.5938e−05	1.5938e−05
	DS_8	1.5938e−05	1.4523–05	1.5938e−05
	DS_9	1.5938e−05	1.3659e−05	1.1758e−05
	DS_{10}	1.5938e−05	1.5938e−05	1.5938e−05
	DS_{11}	1.5938e−05	1.5938e−05	1.5938e−05
	DS_{12}	5.8927e−05	0.0015	0.0025
	DS_{13}	0.0020	1.5938e−05	5.8927e−05
NMI	DS_1	1	1	1
	DS_2	0.0079	0.0079	0.0079
	DS_3	1.5938e−05	1.5938e−05	1.5938e−05
	DS_4	1.5938e−05	1.6874e−05	1.5938e−05
	DS_5	1.5938e−05	1.5938e−05	1.5938e−05
	DS_6	1.5938e−05	1.5938e−05	1.5938e−05
	DS_7	1.4523–05	1.5938e−05	1.5938e−05
	DS_8	1.5938e−05	1.6874e−05	1.5938e−05
	DS_9	1.5938e−05	1.5938e−05	1.5938e−05
	DS_{10}	1.5938e−05	1.5938e−05	1.6874e−05
	DS_{11}	1.6874e−05	1.5938e−05	1.5938e−05
	DS_{12}	4.5938e−05	0.0025	1.5938e−05
	DS_{13}	0.0015	4.5938e−05	3.5938e−05
SI	DS_1	1.5938e−05	1.5938e−05	1.5938e−05
	DS_2	1.5938e−05	1.5938e−05	1.5938e−05
	DS_3	1.5938e−05	1.5938e−05	1.5938e−05
	DS_4	1.6584e−05	1.5938e−05	1.5938e−05
	DS_5	1.5938e−05	1.4521e−05	1.5938e−05
	DS_6	1.5938e−05	1.6584e−05	1.5938e−05
	DS_7	1.5938e−05	1.4521e−05	1.5938e−05
	DS_8	1.6584e−05	1.5938e−05	1.5938e−05
	DS_9	1.5938e−05	1.5938e−05	1.5938e−05
	DS_{10}	1.4521e−05	1.5938e−05	1.6584e−05
	DS_{11}	1.5938e−05	1.4521e−05	1.5938e−05
	DS_{12}	0.0035	1.5938e−05	1.5938e−05
	DS_{13}	3.5938e−05	0.0020	1.0688e−05
DI	DS_1	1.5938e−05	1.5938e−05	1.5938e−05
	DS_2	1.5938e−05	1.5938e−05	1.5938e−05
	DS_3	1.5938e−05	1.5938e−05	1.5938e−05
	DS_4	1.6584e−05	1.5938e−05	1.5938e−05
	DS_5	1.5938e−05	1.4521e−05	1.2546e−05
	DS_6	1.5938e−05	1.6584e−05	1.5938e−05
	DS_7	1.5938e−05	1.4521e−05	1.5938e−05
	DS_8	1.6584e−05	1.5938e−05	1.5938e−05
	DS_9	1.5938e−05	1.5938e−05	1.2546e−05
	DS_{10}	1.4521e−05	1.5938e−05	1.6584e−05
	DS_{11}	1.5938e−05	1.4521e−05	1.5938e−05
	DS_{12}	1.5938e−05	1.3258e−05	1.5688e−05
	DS_{13}	4.5938e−05	1.5938e−05	2.5938e−05
DBI	DS_1	1.5938e−05	1.5938e−05	1.5938e−05
	DS_2	1.5938e−05	1.5938e−05	1.3654e−05
	DS_3	1.5938e−05	1.5938e−05	1.5938e−05
	DS_4	1.6584e−05	1.5938e−05	1.3654e−05
	DS_5	1.5938e−05	1.4521e−05	1.5938e−05
	DS_6	1.5938e−05	1.6584e−05	1.5938e−05
	DS_7	1.5938e−05	1.4521e−05	1.5938e−05
	DS_8	1.6584e−05	1.5938e−05	1.3654e−05
	DS_9	1.3654e−05	1.5938e−05	1.5938e−05
	DS_{10}	1.4521e−05	1.5938e−05	1.6584e−05
	DS_{11}	1.5938e−05	1.4521e−05	1.5938e−05
	DS_{12}	5.1689e−05	1.4589e−05	1.5688e−05
	DS_{13}	1.5938e−05	2.5938e−05	2.8948e−05

Table 4
The computed p-Values based on clustering validity indexes for Wilcoxon's sign rank test to examine M_4 over M_i, where $1 \leq i \leq 3$.

	Dataset	M_1	M_2	M_3
ARI	DS_1	1	1	1
	DS_2	0.0025	0.0025	0.0025
	DS_3	0.0020	0.0020	0.0020
	DS_4	0.0020	0.0020	0.0020
	DS_5	0.0020	0.0020	0.0020
	DS_6	0.0371	0.0440	0.0195
	DS_7	0.0020	0.0020	0.0020
	DS_8	0.0022	0.0020	0.0015
	DS_9	0.0020	0.0020	0.0015
	DS_{10}	0.0020	0.0020	0.0020
	DS_{11}	0.0020	0.0020	0.0020
	DS_{12}	0.0020	0.0015	0.0025
	DS_{13}	0.0020	0.0020	0.0020
NMI	DS_1	1	1	1
	DS_2	0.0025	0.0025	0.0025
	DS_3	0.0020	0.0020	0.0020
	DS_4	0.0020	0.0020	0.0020
	DS_5	0.0020	0.0020	0.0020
	DS_6	0.0025	0.0025	0.0025
	DS_7	0.0020	0.0020	0.0020
	DS_8	0.0025	0.0025	0.0025
	DS_9	0.0025	0.0025	0.0025
	DS_{10}	0.0020	0.0020	0.0020
	DS_{11}	0.0025	0.0025	0.0025
	DS_{12}	0.0020	0.0020	0.0020
	DS_{13}	0.0025	0.0020	0.0015
SI	DS_1	0.0020	0.0020	0.0020
	DS_2	0.0020	0.0020	0.0020
	DS_3	0.0020	0.0020	0.0020
	DS_4	0.0025	0.0020	0.0020
	DS_5	0.0020	0.0020	0.0020
	DS_6	0.0371	0.0840	0.0195
	DS_7	0.0022	0.0020	0.0020
	DS_8	0.0020	0.0025	0.0020
	DS_9	0.0020	0.0015	0.0020
	DS_{10}	0.0020	0.0020	0.0020
	DS_{11}	0.0020	0.0020	0.0020
	DS_{12}	0.0001	0.0015	0.0020
	DS_{13}	0.0015	0.0020	0.0025
DI	DS_1	0.0020	0.0020	0.0020
	DS_2	0.0020	0.0020	0.0020
	DS_3	0.0020	0.0020	0.0020
	DS_4	0.0025	0.0020	0.0020
	DS_5	0.0020	0.0020	0.0020
	DS_6	0.0025	0.0020	0.0020
	DS_7	0.0022	0.0020	0.0020
	DS_8	0.0020	0.0025	0.0020
	DS_9	0.0020	0.0015	0.0020
	DS_{10}	0.0025	0.0020	0.0020
	DS_{11}	0.0022	0.0022	0.0020
	DS_{12}	0.0002	0.0015	0.0020
	DS_{13}	0.0020	0.0005	0.0025
DBI	DS_1	0.0020	0.0020	0.0020
	DS_2	0.0020	0.0020	0.0020
	DS_3	0.0015	0.0020	0.0022
	DS_4	0.0020	0.0020	0.0020
	DS_5	0.0020	0.0020	0.0020
	DS_6	0.0025	0.0020	0.0020
	DS_7	0.0022	0.0020	0.0020
	DS_8	0.0020	0.0025	0.0020
	DS_9	0.0020	0.0015	0.0020
	DS_{10}	0.0020	0.0025	0.0020
	DS_{11}	0.0020	0.0020	0.0020
	DS_{12}	0.0020	0.0005	0.0015
	DS_{13}	0.0002	0.0001	0.0020

in order to conduct Wilcoxon signed, ranksum and signtest. These three experiments help to know whether two dependent data points from populations having same distribution on the obtained values of ARI, NMI, SI, DI and DBI using M_i, where $1 \leq i \leq 4$.

5.4. Results and discussion

Four experiments are performed on all the datasets described in Section 5.1 and the estimated average values of ARI, NMI, SI, DI and DBI using each method, M_i, are reported in Table 1. There is

no doubt that the M_4 outperforms other three methods presented in Section 5.3 because most of the values of the adopted external validity indexes approach 1, which prove the effectiveness of M_4 and it happens since the proposed similarity measure is integrated with traditional FCM. Table 1 also shows the values of used internal validity indexes on all the datasets. All the computed values establish the fact that M_4 is more successful than M_i, where $1 \leq i \leq 3$ because the values produced by M_4 are more precise to ideal values compared to values generated using M_i, where $1 \leq i \leq 3$. The second, third, fourth and fifth rows of Fig. 2 exhibit the cluster-wise data distributions produced by M_1, M_2, M_3 and M_4 respectively on synthetic datasets. Moreover, clustering results on DS_{12} by M_1, M_2, M_3, and M_4 are shown in Figs. 4a, 4b, 4c, and 4d respectively. On the other hand, Figs. 5a, 5b, 5c, and 5d illustrate the clustering outputs of M_1, M_2, M_3, and M_4 on DS_{13} respectively. Some pseudo colors are used to represent each land cover of both the remote sensing datasets. It is clear from Figs. 4 and 5 that all the land covers are accurately identified by M_4. For example, crop land cover is correctly classified by M_4, which is marked by green color in Fig. 5d. On the other hand, all other methods are misclassified crop land cover. These methods recognize crop land cover as forest.

Three non-parametric statistical significance tests namely, Wilcoxon's signed, ranksum and signtest are performed for independent samples at the five percent significance level [34,35]. Four groups based on four methods are formed for each dataset, where each group contains the values of the clustering validity indexes estimated by 10 successive executions of the corresponding method. Table 2 shows the p-values computed by Wilcoxon's signed rank test based on ARI, NMI, SI, DI and DBI separately for comparing two groups namely, M_4 to M_i, where $1 \leq i \leq 3$ at a time. Supposing E_1, E_2, E_3, and E_4 are the median values of each group generated by M_1, M_2, M_3, and M_4 respectively. The null hypothesis attempts to say that no statistically significant variation exists between the median values of two groups namely, M_4 to M_i, where i varies from 1 to 3. It is assumed to be true until statistical evidence nullifies it for an alternative hypothesis. Mathematically, $H_0 : E_1 = E_i \quad vs \quad H_{1i} : E_1 > E_i \quad , where \quad i \in \{2, 3, 4\}$. Most of the p-values noted in Table 2 are less than 5% significance level i.e. 0.05. So, it indicates strong evidence against the null hypothesis, denoting that the better median values of the clustering validity indexes generated by M_4 is statistically significant and it does not happen by chance. Similarly, Tables 3 and 4 report the p-values obtained by the ranksum and sign test. The most of the p-values of Tables 3 and 4 are less than 0.5. So, we can reject the null hypothesis for 5% confidence level. Therefore, the proposed method M_4 outperforms M_i based on all the results.

6. Conclusion

This work proposes a novel similarity measure on $\mathbb{R}_+^n$ by considering Jeffreys-divergence. Various properties of the proposed similarity measure are discussed. Traditional FCM algorithm is revised by replacing the Euclidean distance with the help of the proposed similarity measure. A theoretical analysis of the modified FCM is addressed by furnishing the detail proof of convergence. The modified FCM algorithm guarantees to a local minima. We validate our claim through detailed experimental results and statistical analysis on some synthetic and real-world datasets including two sets of satellite images. The study of data complexity metrics is a promising area of research in the field of clustering because the performance depends on data points of a dataset. It deserves further study.

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.asoc.2019.106016.

CRediT authorship contribution statement

Ayan Seal: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **Aditya Karlekar:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **Ondrej Krejcar:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **Consuelo Gonzalo-Martin:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing.

Acknowledgments

This work is partially supported by the project of Grant Agency of Excellence, University of Hradec Kralove, Faculty of Informatics and Management, Czech Republic (under ID: UHK-FIM-GE-2019) and project of the Ministry of Education, Youth and Sports of Czech Republic (project ERDF no. CZ.02.1.01/0.0/0.0/18_069/0010054).

References

[1] S. Yin, Z. Huang, Performance monitoring for vehicle suspension system via fuzzy positivistic c-means clustering based on accelerometer measurements, IEEE/ASME Trans. Mechatronics 20 (5) (2014) 2613–2620.
[2] W. Bi, M. Cai, M. Liu, G. Li, A big data clustering algorithm for mitigating the risk of customer churn, IEEE Trans. Ind. Inf. 12 (3) (2016) 1270–1281.
[3] S. Yin, H. Gao, J. Qiu, O. Kaynak, Fault detection for nonlinear process with deterministic disturbances: a just-in-time learning based data driven method, IEEE Trans. Cybern. 47 (11) (2016) 3649–3657.
[4] X. Cao, X. Wei, Y. Han, D. Lin, Robust face clustering via tensor decomposition, IEEE Trans. Cybern. 45 (11) (2014) 2546–2557.
[5] X. Pei, T. Wu, C. Chen, Automated graph regularized projective nonnegative matrix factorization for document clustering, IEEE Trans. Cybern. 44 (10) (2014) 1821–1831.
[6] S. Yin, X. Zhu, J. Qiu, H. Gao, State estimation in nonlinear system using sequential evolutionary filter, IEEE Trans. Ind. Electron. 63 (6) (2016) 3786–3794.
[7] J. MacQueen, et al., Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, Oakland, CA, USA, 1967, pp. 281–297.
[8] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., A density-based algorithm for discovering clusters in large spatial databases with noise, in: Kdd, vol. 9, 1996, pp. 226–231.
[9] B.J. Frey, D. Dueck, Clustering by passing messages between data points, Science 315 (5814) (2007) 972–976.
[10] J. Shi, J. Malik, Normalized cuts and image segmentation, Departmental Papers (CIS), 2000, p. 107.
[11] A. Banerjee, S. Merugu, I.S. Dhillon, J. Ghosh, Clustering with Bregman divergences, J. Mach. Learn. Res. 6 (Oct) (2005) 1705–1749.
[12] A. Saha, S. Das, Geometric divergence based fuzzy clustering with strong resilience to noise features, Pattern Recognit. Lett. 79 (2016) 60–67.
[13] S. Chakraborty, S. Das, K- means clustering with a new divergence-based distance metric: convergence and performance analysis, Pattern Recognit. Lett. 100 (2017) 67–73.

[14] L. Legrand, E. Grivel, Jeffrey's divergence between moving-average models that are real or complex, noise-free or disturbed by additive white noises, Signal Process. 131 (2017) 350–363.
[15] F. Nielsen, R. Nock, Total Jensen divergences: definition, properties and clustering, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015, pp. 2016–2020.
[16] F. Nielsen, R. Nock, S.-i. Amari, On clustering histograms with k-means by using mixed α-divergences, Entropy 16 (6) (2014) 3273–3301.
[17] R. Nock, F. Nielsen, S.-i. Amari, On conformal divergences and their population minimizers, IEEE Trans. Inform. Theory 62 (1) (2015) 527–538.
[18] M. Das Gupta, S. Srinivasa, M. Antony, et al., KL divergence based agglomerative clustering for automated vitiligo grading, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2700–2709.
[19] A. Notsu, O. Komori, S. Eguchi, Spontaneous clustering via minimum gamma-divergence, Neural Comput. 26 (2) (2014) 421–448.
[20] J.C. Dunn, A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters, Taylor & Francis, 1973.
[21] W. Peizhuang, Pattern recognition with fuzzy objective function algorithms (James C. Bezdek), SIAM Rev. 25 (3) (1983) 442.
[22] L. Groll, J. Jakel, A new convergence proof of fuzzy c-means, IEEE Trans. Fuzzy Syst. 13 (5) (2005) 717–720.
[23] F. Hoppner, F. Klawonn, A contribution to convergence theory of fuzzy c-means and derivatives, IEEE Trans. Fuzzy Syst. 11 (5) (2003) 682–694.
[24] N.R. Pal, J.C. Bezdek, R.J. Hathaway, Sequential competitive learning and the fuzzy c-means clustering algorithms, Neural Netw. 9 (5) (1996) 787–796.
[25] W. Wei, J.M. Mendel, Optimality tests for the fuzzy c-means algorithm, Pattern Recognit. 27 (11) (1994) 1567–1573.
[26] D. Bertsekas, Nonlinear Programming, Athena Scientific, Belmont (MA), 1995.
[27] D. Dheeru, E. Karra Taniskidou, UCI machine learning repository, 2017, http://archive.ics.uci.edu/ml.
[28] J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. Garcí a, L. Sánchez, F. Herrera, KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework, J. Mult.-Valued Logic Soft Comput. 17 (2011).
[29] A. Seal, A. Garcia-Pedrero, D. Bhattacharjee, M. Nasipuri, M. Lillo-Saavedra, E. Menasalvas, C. Gonzalo-Martin, Multi-scale RoIs selection for classifying multi-spectral images, Multidimens. Syst. Signal Process. (2019) 1–25.
[30] U. Maulik, S. Bandyopadhyay, Performance evaluation of some clustering algorithms and validity indices, IEEE Trans. Pattern Anal. Mach. Intell. 24 (12) (2002) 1650–1654.
[31] L. Hubert, P. Arabie, Comparing partitions, J. Classif. 2 (1) (1985).
[32] N.X. Vinh, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance, J. Mach. Learn. Res. 11 (Oct) (2010).
[33] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.
[34] A. Karlekar, A. Seal, O. Krejcar, C. Gonzalo-Martin, Fuzzy K-means using non-linear S-distance, IEEE Access 7 (2019) 55121–55131.
[35] K.K. Sharma, A. Seal, Modeling uncertain data using Monte Carlo integration method for clustering, Expert Syst. Appl. 137 (2019) 100–116.