Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Robust denoising FCM clustering via $L_{2,1}$ NMF and local constraint

Xuezhen Fan[1,3] | **Xiangli Li**[1,2] | **Xiyan Lu**[1,3]

[1]School of Mathematics and Computing Science, Guilin University of Electronic Technology, Guilin, China

[2]Guangxi Colleges and Universities Key Laboratory of Data Analysis and Computation, Guilin University of Electronic Technology, Guilin, China

[3]Guangxi Applied Mathematics Center, Guilin University of Electronic Technology, Guilin, China

Correspondence
Corresponding author Xiangli Li, School of Mathematics and Computing Science, Guilin University of Electronic Technology, Guilin, China.
Email: lixiangli@guet.edu.cn

Present address
Guilin University of Electronic Technology, Guilin, Guangxi Zhuang Autonomous Region, China

Abstract

The Fuzzy C-Means (FCM) algorithm is widely used in data mining and machine learning. However, the sensitivity of FCM to the initial value and noise inevitably leads to the decline of the clustering effect. In this paper, a new improved fuzzy clustering algorithm is proposed— Robust denoising FCM clustering via $L_{2,1}$ NMF and local constraint (RFCM-$L_{2,1}$NMF). Firstly, RFCM-$L_{2,1}$NMF combines the $L_{2,1}$NMF that has noise residual estimation with FCM, using the robustness and noise constraint terms of the $L_{2,1}$NMF to attenuate the effect of noise on data clustering. Secondly, RFCM-$L_{2,1}$NMF uses the low-dimensional representation of $L_{2,1}$NMF as the initial value of FCM, which reduces the defects of FCM caused by the initial value to a certain extent, and makes the clustering effect more stable. Furthermore, since the low-dimensional representation of $L_{2,1}$NMF is the hub connecting $L_{2,1}$NMF and FCM, to obtain a more accurate low-dimensional representation, we construct a new local constraint term in this paper. Finally, experiments on data sets validate that RFCM-$L_{2,1}$NMF is superior compared to other state-of-the-art methods.

KEYWORDS
fuzzy C-means, $L_{2,1}$ Nonnegative matrix factorization, noise residual, local constraint

1 | FIRST LEVEL HEAD

Data clustering is one of the basic topics in machine learning. Its purpose is to divide data samples into different clusters according to a certain criterion, so that the data samples of the same cluster have high similarity, and the data of different clusters difference is as large as possible. Cluster analysis is also widely used in data mining[1], pattern recognition[2], image processing[3] and many other fields.

The fuzzy clustering algorithm[4] is a classical partition-based clustering algorithm. Fuzzy clustering algorithm introduces the concept of fuzzy, to establish the uncertainty description of the sample to the category, and expand the value range of the membership degree. At the same time, fuzzy clustering algorithm uses the degree of membership to determine that each sample point belongs to a certain cluster, so that it has a better clustering effect and data expression ability, can more objectively reflect the objective world, and make the classification more realistic, so fuzzy clustering has become a research hotspot of cluster analysis. Fuzzy clustering algorithms mainly include transitive closure methods based on fuzzy equivalence[5], methods based on similarity relations and fuzzy relations[6], and maximum tree methods based on fuzzy graph theory[7] etc. But these methods have been gradually reduced in practical applications and research due to their high computational complexity. Because of good robustness and flexibility, the most widely used in practical applications is the fuzzy clustering algorithm based on the objective function[4], that is, the fuzzy C-means clustering algorithm (FCM).

However, the traditional FCM algorithm still has some defects, such as being very sensitive to the initial value and noise, slow convergence speed and a large amount of calculation. To solve these problems, experts and scholars have proposed some variant FCM algorithms. Hathaway and Hu[8] designed a density-weighted fuzzy C-means clustering (DWFCM) to improve convergence speed by simplifying a larger data set into a smaller weighted data set. Hung et al.[9] refined the initial value of the FCM algorithm and proposed a psFCM algorithm. Based on the FCM method, Gao et al.[10] combined with relative entropy,

Abbreviations: FCM, Fuzzy C-Means; NMF, Nonnegative matrix factorization; RFCM, Robust denoising FCM.

and proposed a new method to intelligently consider noise-adaptive FCM and its extended version (adaptive-REFCM). Liu[11] proposed a new FCM clustering algorithm based on local density. At the same time, some improved FCM algorithms were designed to process large-scale clustering problems. The gradient-based fuzzy C-means (GBFCM)[12] used the gradient decrease to improve convergence speed and stability. Given that the high dimensionality of features may lead to a high-complexity and low-stability clustering performance, Havens et al.[13] presented LFCM and rseFCM algorithms for very large data, which reduce the complexity of clustering through nonlinear clustering with kernel techniques and relaxing convergence conditions. In addition, Li et al.[14] proposed a fuzzy C-means clustering algorithm with different attributes, and obtained a new parameter selection rule. Krinidis et al.[15] used local spatial information and gray information in a new fuzzy way. Fuzzy local information C-means algorithm (FLICM) was proposed by fusion together. Recently, based on the triangle inequality, Zhou et al.[16] proposed a new membership fuzzy C-means clustering algorithm (MSFCM). Wang et al.[17] proposed a new Residual-driven FCM with weighted L_2-norm fidelity (WRFCM) algorithm, which is based on residual estimation and obtains a weighted L_2-norm fidelity term by weighting the mixed noise distribution, thereby reducing the impact of noise on clustering. In the era of information technology, the explosion of information has made the structure of data more complex. However, the existing FCM clustering algorithm lacks a direct connection with the structural characteristics of the data set, so its effectiveness in processing clustering problems is far from satisfactory.

Lee et al.[18] proposed non-negative matrix factorization in 1999, whose purpose is to obtain two low-rank non-negative factor matrices and make their product close to the original data matrix. Non-negative matrix factorization (NMF) is widely used due to its well-interpreted and clear physical meaning[19,20]. Many representative variants and extensions of NMF have subsequently been proposed to solve different problems. For example, since many data contain noise and outliers, Kong et al.[21] proposed a robust formulation of NMF using $L_{2,1}$ norm loss function($L_{2,1}$ NMF), and derive a computational algorithm with rigorous convergence analysis. Hoyer et al.[22] proposed non-negative sparse coding (NNSC), this method directly used the l_1-norm on the coding matrix to enhance the sparsity of the decomposition results. Cai et al.[23] proposed the graph regular non-negative matrix factorization (GNMF), which used the graph Laplace matrix to preserve the geometric structure of the data and significantly improved the effect of clustering. Recently, Ye et al.[24] proposed ensemble clustering based on non-negative matrix factorization without using prior information. Tong et al.[25] proposed non-negative matrix factorization with local constraints to improve action recognition accuracy. In addition, in[26], Tao et al. presented image clustering methods based on non-negative matrix factorization and fuzzy C-means, etc.

Inspired by the above work, we propose a new algorithm in this paper, Robust denoising FCM clustering via $L_{2,1}$ Nonnegative matrix factorization (RFCM-$L_{2,1}$NMF). RFCM-$L_{2,1}$NMF absorbs the advantages of both $L_{2,1}$NMF and FCM. Specifically, our algorithm uses fuzzy clustering in a low-dimensional subspace of the original data, which avoids the FCM being affected to some extent by the initial value. At the same time, to strengthen the connection with the structural features of the data set, we construct a new local constraint term to make the low-dimensional representation more accurate. In addition, a control noise term is added to the new objective function to reduce the effect of noise on clustering.

The rest of the paper is structured as follows. Section 2 introduces the related work of FCM and $L_{2,1}$NMF. In Section 3, RFCM-$L_{2,1}$NMF are proposed. In Section 4, some comparative experiments are done to verify the effectiveness of the proposed algorithm. The conclusion is given in the last section.

2 | RELATED WORKS

2.1 | NMF

Given a non-negative data matrix $X = [x_1, x_2, \cdots, x_n] \in R^{p\times n}$, and $x_i \in R^p$ is the i–th column of X, which represents a data point. The standard NMF decomposes X as a product of the basis matrix $W \in \mathbb{R}^{p\times c}$ and the coefficient matrix $H \in \mathbb{R}^{c\times n}$, where H is the representation of the original data in a low-dimensional space, c is the reduced dimension. NMF is widely used in machine learning fields and data mining. Since it uses Euclidean distance, i.e. the error of each data point is squared into the objective function, which makes it prone to outliers. Kong et al.[21] proposed a robust formulation of NMF using $l_{2,1}$ norm loss function.

$$\begin{aligned} &\min_{W,H} \quad \|X - WH\|_{2,1}, \\ &s.t. \quad W \geq 0, H \geq 0 \end{aligned} \tag{1}$$

where $\|\cdot\|_F$ represents the Frobenius norm.

From the literature[21], the iterative update formula of (1) can be obtained as follows:

$$W_{ij} \leftarrow W_{ij} \frac{(XDH^T)_{ij}}{(WHDH^T)_{ij}}, \tag{2}$$

$$H_{ij} \leftarrow H_{ij} \frac{(W^T XD)_{ij}}{(W^T WHD)_{ij}} . \tag{3}$$

where D is a diagonal matrix with the diagonal elements given by

$$D_{ii} = 1/\sqrt{\sum_{j=1}^{p} (X - WH)_{ji}^2} = \frac{1}{\|x_i - Wh_i\|} . \tag{4}$$

.

2.2 | FCM

FCM clustering partition X into c clusters with the cluster centers $V = [v_1, v_2, \cdots, v_c] \in \mathbb{R}^{p\times c}$ and the membership degree matrix $U = \left[u_{ij}\right] \in \mathbb{R}^{c\times n}$. For a given fuzziness weighting exponent $m > 1$, V and U are solved iteratively according to the following optimization problem

FCM clustering partition X into c clusters with the cluster centers $V = [v_1, v_2, \cdots, v_c] \in \mathbb{R}^{p\times c}$ and the membership degree matrix $U = \left[u_{ij}\right] \in \mathbb{R}^{c\times n}$. For a given fuzziness weighting exponent $m > 1$, V and U are solved iteratively according to the following optimization problem

$$\begin{aligned} \min_{U,V} \ & J_m = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m \|x_k - v_i\|^2 \\ & \text{s.t.} \sum_{i=1}^{c} u_{ik} = 1, \end{aligned} \tag{5}$$

where $v_{.j}$ represents the j–th column row of V, $u_{i.}$ represents the i–th row of U.

The FCM scheme usually initializes U as $U^{(0)}$ and updates V and U alternatively by

$$v_i = \frac{\sum_{j=1}^{n} \left(u_{ij}\right)^m x_j}{\sum_{j=1}^{n} \left(u_{ij}\right)^m},$$

$$u_{ij} = \left[\sum_{k=1}^{c} \left(\frac{\|x_j - v_i\|}{\|x_j - v_k\|}\right)^{\frac{2}{m-1}}\right]^{-1} .$$

until convergence in achieved.

In the process of clustering, noise is a problem that cannot be ignored. If the noise distribution can be made closer to the Gaussian distribution, the noise can be characterized using the l_2 norm, which means that the noise can be estimated more accurately. In[21], it is proposed to assign an appropriate weight to each noise element, which forms a weighted residual that almost obeys a Gaussian distribution. As follows,

$$\|W_l \odot R_l\|_2^2 = \sum_{j=1}^{k} \left|w_{jl} r_{jl}\right|^2. \tag{6}$$

where $\odot$ performs element-by-element multiplication, each element w_{jl} is assigned to a location (j, l), and $w_{jl} = e^{-\xi r_{jl}^2}$, where ξ is a positive parameter, which aims to control the decreasing rate of w_{jl}.

Moreover, Zhou et al.[16] introduced the triangle inequality in clustering based on FCM and gave a new geometric explanation, and proposed a new membership scaling scheme. The basic idea of Zhou is to use the triangle inequality to filter out those samples whose nearest cluster centers do not change in the next iteration. By using a new scaling scheme, the effect of the in-cluster samples is enhanced, and the relationship of samples out-of-cluster is weakened.

2.3 | Entropy-like divergence kernel function

To solve the drawbacks of Euclidean distance in dealing with noise, and inspired by information divergence and S-divergence, Wu et al.[27] constructed an entropy-like divergence kernel by combining the Jensen-Shannon/Bregman divergence with a convex function $\varphi(x) = -x \ln x$. Setting $d_e(x, y) = \frac{x \ln x + y \ln y}{2} - \frac{x+y}{2} \ln(\frac{x+y}{2})$ and entropy-like divergence kernel is defined as follows:

$$K_{\mathrm{e}}(x, y) = e^{-\frac{d_e(x,y)}{2\delta^2}}, \tag{7}$$

where δ is the scale parameter of entropy-like divergence kernel.

Compared with Euclidean distance, entropy-like divergence can suppress noise better. For square Euclidean distance, minimizing $\sum_{j=1}^{n} \|x_j - z\|^2$ for z, a sample mean is obtained as follows.

$$z = \left(\sum_{j=1}^{n} x_j \right) /n, \tag{8}$$

For entropy-like divergence, minimization $\sum_{j=1}^{n} (\frac{x_j \ln x_j + z \ln z}{2} - \frac{x_j+z}{2} \ln(\frac{x_j+z}{2}))$ for z and the sample mean is obtained as follows.

$$z = \exp\left(\frac{1}{n} \cdot \ln(\frac{x_j + z}{2}) \right). \tag{9}$$

The artificial data set $\{31.1, 31.6, 31.9, 32, 32.2, 32.4, 32.5, 32.8, 33.5, 34\}$ is given and used to test the robustness of entropy-like divergence. By solving Eq. (8) and Eq. (9), the estimated value of z is 32.4. Set the noise point to 17. In this case, Eq. (8) can obtain its estimate as z = 31, and Eq. (9) can obtain its estimate as $z = 31.1002$, which means that they are both corrupted by noise points. But equation Eq. (8) yields values that are outside the original data range, while Eq. (9) yields values that do not. So we know that both the squared Euclidean distance and the class entropy divergence are affected by noise points, but compared with squared Euclidean distance, the center value obtained by entropy-like divergence is closer to the estimated center value of 32.4, which means that entropy-like divergence can better able to suppress noise than Euclidean distance.

3 | THE PROPOSED ALGORITHM

3.1 | RFCM-$L_{2,1}$NMF

NMF is essentially a dimensionality reduction tool. Our algorithm retains the advantages of NMF and combines the low-dimensional representation of NMF with FCM, which weakens the influence of initial values on FCM to a certain extent. At the same time, to reduce the influence of noise points and outliers on clustering, we add noise sparse terms and local constraints to the objective function.

3.1.1 | Objective function of RFCM-$L_{2,1}$NMF

Since NMF is a popular and effective dimensionality reduction method, we absorb the advantages of NMF to make up for the large-scale calculation problem caused by FCM clustering on high-dimensional data sets. To obtain a more accurate low-dimensional representation, we consider adding the local constraint term to the objective function, namely

$$\sum_{i=1}^{c} \sum_{j=1}^{n} h_{ij} k_{ij} = tr(H^T K), \tag{10}$$

where $K = \left[k_{ij}\right] \in \mathbb{R}^{c\times n}$ measures the spatial relationship between x_j and w_i. For further denoising, our K adopts Entropy-like divergence kernel function from the previous section,as follows,

$$\begin{aligned} &K = 1 - R_{\mathrm{e}}(x_j, w_i), \\ &R_{\mathrm{e}}(x, w) = e^{-\frac{d_e(x,w)}{2\delta^2}}, \\ &d_e(x, w) = \tfrac{x\ln x + w\ln w}{2} - \tfrac{x+w}{2}\ln(\tfrac{x+w}{2}). \end{aligned} \tag{11}$$

RFCM-$L_{2,1}$NMF uses NMF for dimensionality reduction, and uses fuzzy clustering on low-dimensional subspaces, which effectively avoids large-scale calculation problems caused by direct implementation of fuzzy clustering on high-dimensional data sets. At the same time, to reduce the influence of noise, a noise constraint term is added to the objective function, and the minimized cost function is as follows:

$$\begin{aligned} \min_{W,H,S,U,V,K} \quad & \|X - WH - S\|_{2,1} \\ & + \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m}\|h_j - v_i\|^2 + \alpha\,\|G \odot S\|_F^2 \\ & + \beta tr(H^T K), \\ s.t.\ \ 0 \le u_{ij} \le 1,\ & \sum_{i=1}^{c} u_{ik} = 1, \quad W, H, U, V \ge 0 \end{aligned} \tag{12}$$

where $G = e^{-\xi S}$, the first term is the $L_{2,1}$ NMF with noise residual, the second term is the FCM on the low-dimensional subspace obtained by NMF, the third term is the noise constraint term, and the fourth term is the local constraint term.

3.1.2 | Optimization Algorithm

Obviously, optimization problem (12) is non-convex, and solving all variables is NP hard. To solve this problem, a method similar to one in [18]. Next, the optimal solution is solved by alternating iterative optimization.

(1) Firstly, fixed W, H, S, K and V, according to FCM, we can get the iterative formula of U as

$$u_{ij} = \left(\sum_{k=1}^{c}\left(\frac{\|h_j - v_i\|}{\|h_j - v_k\|}\right)^{\frac{2}{m-1}}\right)^{-1} \tag{13}$$

(2) Then, When W, S, K, U and V are fixed, let $\hat{X} = X - S$, then (12) can be reduced to

$$\begin{aligned} \min_{H\ge 0} J_m = & \left\|\hat{X} - WH\right\|_{2,1} + \sum_{i=1}^{n}\left\|\left(h_i\mathbf{1}^T - V\right)\Lambda_i^{1/2}\right\|_F^2 \\ & + \beta tr(H^T K) \end{aligned} \tag{14}$$

where $\mathbf{1}$ is the all 1-column vector of dimension, $\Lambda_i = \begin{bmatrix} u_{1i} & & & \\ & u_{2i} & & \\ & & \ddots & \\ & & & u_{ci} \end{bmatrix}$. Let ϕ_{ij} be the Lagrange multiplier of constraint h_{ij}, and $\Phi = \left[\phi_{ij}\right]$; The Lagrangian function corresponding to the above formula is

$$\begin{aligned} L = & \left\|\hat{X} - WH\right\|_{2,1} + \sum_{i=1}^{n}\left\|\left(h_i\mathbf{1}^T - V\right)\Lambda_i^{1/2}\right\|_F^2 \\ & + \beta tr(H^T K) + tr(\Phi H^T). \end{aligned} \tag{15}$$

Taking the partial derivative of (15) with respect to H is calculated, and then the formula is obtained as follows:

$$\begin{aligned}\frac{\partial L}{\partial H} &= -W^T\hat{X}D + W^TWHD - V(U^m) \\ &\quad + H \odot [\mathbf{1} * sum(U^m)] + \beta K + \Phi,\end{aligned} \tag{16}$$

According to the Karush – Kuhn – Tucker (KKT) condition, the complementary slackness condition $\phi_{ij}h_{ij} = 0$ and zero gradient condition $\frac{\partial L}{\partial H}$=0, we get

$$\begin{aligned}&\left(-W^T\hat{X}D + W^TWHD - V(U^m)\right)_{ij}h_{ij} \\ &+(H \odot [\mathbf{1} * sum(U^m)] + \beta K + \Phi)_{ij}h_{ij} = 0,\end{aligned} \tag{17}$$

Therefore,update the variables H, which can be given as follows

$$H_{ij} \leftarrow H_{ij}\frac{\left(W^T\hat{X}D + V\left(U^m\right)\right)_{ij}}{\left(W^TWHD + H \odot (\mathbf{1} * \operatorname{sum}\left(U^m\right)) + \beta K\right)_{ij}}. \tag{18}$$

where D is a diagonal matrix with the diagonal elements given by
$D_{ii} = 1/\sqrt{\sum_{j=1}^{m}(\hat{X} - WH)_{ji}^2} = \frac{1}{\left\|\hat{x}_i - Wh_i\right\|}$.

(3) Next, when S, K, U, V and H are fixed, we update the variables W by $L_{2,1}$NMF.

$$W_{ij} \leftarrow W_{ij}\frac{(\hat{X}DH^T)_{ij}}{(WHDH^T + \beta Q)_{ij}}. \tag{19}$$

(4) Similar to[16], we also use filtering techniques

Let $\delta_i = \left\|v_i^{(t+1)} - v_i^{(t)}\right\|$ be displacement of the cluster center $v_i^{(t)}(1 \le i \le c)$. For any h_j, the distances between h_j and the current cluster centers $V^{(t)}$ are $l_{ij}^{(t)} = \left\|h_j - v_i^{(t)}\right\|$ $(1 \le i \le c)$ These are rearranged in ascending order and denoted as $L_j^{(1)}, L_j^{(2)}, \cdots, L_j^{(c)}$, that is, $L_j^{(1)} \le L_j^{(2)} \le \cdots \le L_j^{(c)}$,its nearest cluster center $v_{I_j^*}^{(t)}\left(I_j^* = \arg\min_{1\le i\le c}\left\{l_{i,j}^{(t)}\right\}\right)$ does not change after another update, if

$$L_j^{(2)} - \max_{1\le i\le c}\delta_i \ge L_j^{(1)} + \delta_{I_j^*} \tag{20}$$

That is, $\arg\min_{1\le i\le c}\left\{l_{i,j}^{(t+1)}\right\} = I_j^*$ holds in this case. In addition, the label $H_{Q_t} = \left\{h_j \mid j \in Q_t\right\}$ is a sample filtered through inequality (20).

(5) Membership degree scaling scheme

For H_{Q_t}, the membership degrees to the cluster centers $V^{(t)}$ are the vector $u_j^{(t)} = \left(u_{1,j}, \cdots, u_{c,j}\right)^\top$. If $i = I_j^*$, $u_{I_j^*,j}^{(t)}$ is increased to $u_{I_j^*,j}^{(t+1)}$ in the next iteration, otherwise, the value of $u_{i,j}^{(t)}$ is decreased. Among them, we scale the value of the degree of membership similar to the literature[16]. The update rule of the membership degree is as follows:

$$u_{i,j}^{(t+1)} = \begin{cases} M_j, & j \in Q_t, i = I_j^* \\ \beta_j u_{i,j}^{(t)}, & j \in Q_t, i \ne I_j^* \\ u_{i,j}^{(t)}, & j \notin Q_t, 1 \le i \le c, \end{cases} \tag{21}$$

where

$$M_j = \left[1 + (c-1)\left(\frac{L_j^{(1)}}{L_j^{(c)}}\right)^{\frac{2}{m-1}}\right]^{-1}, \quad \beta_j = \frac{1 - M_j}{1 - u_{I_j,j}^{(t)}} \tag{22}$$

.

(6) When W, H, S, K and U are fixed, according to FCM, we can get the iterative formula of V as

$$v_i = \frac{\sum_{j=1}^{n}\left(u_{ij}\right)^m h_j}{\sum_{j=1}^{n}\left(u_{ij}\right)^m}. \tag{23}$$

(7) Finally,When W, H, K, U and V are fixed,let $\tilde{X} = X - WH$, then the problem (12) can be reduced to

$$\min_{S\geq 0} J_m = \left\|\tilde{X} - S\right\|_{2,1} + \alpha \left\|G \odot S\right\|_F^2, \tag{24}$$

Let ψ_{ij} be the Lagrange multiplier of constraint s_{ij}, and $\Psi = \left[\psi_{ij}\right]$; The Lagrangian function corresponding to the above formula is

$$L = \left\|\tilde{X} - S\right\|_{2,1} + \alpha \left\|G \odot S\right\|_F^2 + \mathrm{tr}(\Psi S), \tag{25}$$

Taking the partial derivative of (25) with respect to H is calculated, and then the formula is obtained as follows:

$$\frac{\partial L}{\partial S} = -(X - WH - S)\mathrm{D1} + \alpha N + \Psi, \tag{26}$$

According to the Karush – Kuhn – Tucker (KKT) condition, the complementary slackness condition $\psi_{ij}s_{ij} = 0$ and zero gradient condition $\frac{\partial L}{\partial S}$=0, we get

$$(-(X - WH - S)\mathrm{D1} + \alpha N)_{ij}s_{ij} + \psi_{ij}s_{ij} = 0, \tag{27}$$

$$S_{ij} \leftarrow S_{ij}\frac{(XD1)_{ij}}{((WH + S)D1 + \alpha N)_{ij}}, \tag{28}$$

where D1 is a diagonal matrix with the diagonal elements given by
$D1_{ii} = 1/\sqrt{\sum_{j=1}^{p}(\tilde{X} - S)_{ji}^2} = \frac{1}{\|\tilde{x} - s_i\|}$.

Algorithm 1 RFCM-$L_{2,1}$NMF

Input: Dataset $X = [x_1, x_2, \cdots, x_n]$, initial membership degree matrix $W^{(0)}$, $H^{(0)}$, $U^{(0)}$, $K^{(0)}$, $S^{(0)}$, cluster number c, fuzzy exponent m, and convergence threshold ε

Output: Membership degree matrix U and cluster center matrix V;

1: Compute the cluster center $V^{(1)}$ by the initial membership degree matrix $U^{(0)}$ according to 23, Set $t := 1$;
2: Compute $U^{(t)}$ by (13);
3: Compute $H^{(t)}$ by (18);
4: Compute $W^{(t)}$ by (19);
5: Compute $K^{(t)}$ by (11);
6: Compute $S^{(t)}$ by (28);
7: Compute $\bar{V}^{(t)}$ by (23);
8: Compute $\delta_i = \left\|\bar{v}_i^{(t)} - v_i^{(t-1)}\right\|$;
9: Filter out the subset H_{Q_t}according to (20);
10: Update $U^{(t+1)}$ with the new scheme according to (21);
11: Update $H^{(t+1)}$ by (18);
12: Update $W^{(t+1)}$ by(19);
13: Compute $K^{(t+1)}$ by (11);
14: Compute $S^{(t+1)}$ by (28);
15: Update $V^{(t+1)}$ by(23);
16: **if** $\left\|V^{(t+1)} - V^{(t)}\right\| \leq \varepsilon$ **then**
return $U = U^{(t+1)}$, $V = V^{(t+1)}$
else
Set $t := t + 1$, Go to Step 2
end if

4 | EXPERIMENT

In this section, we conduct experiments on seven benchmark data sets to evaluate the performance of RFCM-$L_{2,1}$ NMF . First of all, to illustrate the effectiveness of the proposed RFCM-$L_{2,1}$ NMF method in clustering tasks, we make a comparison between RFCM-$L_{2,1}$ NMF and four related methods, which are $L_{2,1}$ NMF[21], FCM[4], LFCM[13], MSFCM[16]. $L_{2,1}$ NMF is a robust NMF

TABLE 1 Description of these datasets.

Datasets	Samples	Dimensions	Classes
Iris	150	4	3
Wine	178	13	3
satimage	2310	36	6
Yale	165	1024	15
USPS	9298	256	10

†Example for a first table footnote.

clustering algorithm. FCM is a classic fuzzy clustering algorithm. LFCM is an FCM algorithm for big data. MSFCM has a great correlation with the two algorithms proposed in this article.

4.1 | Datasets and Evaluation Measures

In our experiments, five benchmark data sets are used, i.e., satimage, Wine, Iris, USPS and Yale. Yale is a dataset of face images, the USPS data set is from[28], while the other three data sets are from the UCI machine learning repository[16]. The detailed information of the benchmark data sets is given in Table 1.

To quantitatively evaluate the clustering results, three widely used evaluation measures are adopted, namely, the overall F-measure for the entire data set ($\mathbf{F}^*$)[29], normalized mutual information (**NMI**)[30] and adjusted Rand index (**ARI**)[31]. Note that large values of $\mathbf{F}^*$, **NMI** and **ARI** indicate better clustering results.

Let n be the total number of samples, $\{\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_c\}$ be the partition of the ground truth, and $\left\{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2, \cdots, \hat{\mathcal{C}}_{\hat{c}}\right\}$ be the partition by an algorithm. Denote that $\hat{n}_i = \left|\hat{\mathcal{C}}_i\right|$ is the number of samples in $\hat{\mathcal{C}}_i$, $n_l = \left|\mathcal{C}_l\right|$ is the number of samples in $\mathcal{C}_l$, and $n_i^l = \left|\mathcal{C}_l \cap \hat{\mathcal{C}}_i\right|$ is the number of the common objects in $\mathcal{C}_l$ and $\hat{\mathcal{C}}_i$, where $1 \le i \le \hat{c}$ and $1 \le l \le c$.

Then the measure $F(l,i) = \frac{2n_i^l}{n_l + \hat{n}_i}$ is the harmonic mean of the precision and recall of $\mathcal{C}_l$ and its potential prediction $\hat{\mathcal{C}}_i$. Therefore, the overall F-measure $\mathbf{F}^*$, **NMI** , and **ARI** are defined as the following equations:

$$\mathbf{F}^* = \sum_{l=1}^{c} \frac{n_l}{n} \max\{F(l,i) \mid i = 1, \cdots, \hat{c}\}, \tag{29}$$

$$\mathbf{ARI} = \frac{\sum_{i=1}^{\hat{c}} \sum_{l=1}^{c} \binom{n_i^l}{2} - K}{\frac{1}{2}\left(\sum_{i=1}^{\hat{c}} \binom{t_i}{2} + \sum_{l=1}^{c} \binom{s_l}{2}\right) - K}, \tag{30}$$

where $K = \sum_{i=1}^{\hat{c}} \binom{t_i}{2} \sum_{l=1}^{c} \binom{s_l}{2} / \binom{n}{2}$.

$$\mathbf{NMI} = \frac{\sum_{i=1}^{\hat{c}} \sum_{l=1}^{c} n_i^l \log\left(\frac{n \cdot n_i^l}{\hat{n}_i \cdot n_l}\right)}{\sqrt{\left(\sum_{i=1}^{\hat{c}} \hat{n}_i \log\left(\frac{\hat{n}_i}{n}\right)\right)\left(\sum_{l=1}^{c} n_l \log\left(\frac{n_l}{n}\right)\right)}}. \tag{31}$$

4.2 | Parameter analysis

In this subsection, for fair comparisons, parameters in related methods are the same as original papers. For all datasets, δ is set as 1. Meanwhile, in RFCM-$L_{2,1}$NMF, the parameters that affect the experimental result are α, β ,m and ξ, where α is used to control noise constraint term, β is used to control local constraint term and ξ is a positive parameter used to control the rate of change of the noise weight. The fuzziness parameter m is a key parameter that can affect the result of the FCM clustering. Bezdek and Hathaway[30] analyzed the convergence of the FCM algorithm and found that the value of m was related to the number of samples n, and the value of m is recommended to be greater than $\frac{n}{n-2}$.

In order to evaluate the sensitivity of α, β and ξ on the performance of RFCM-$L_{2,1}$NMF, we tune α, β and ξ in the range of $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}\}$, the parameters m is selected from $\{1.2, 1.4, 1.6, 1.8, 2, 2.1\}$. The parameter selection will be analyzed from the experimental results.

we discuss the selection of the parameters α, β, m and ξ for RFCM-$L_{2,1}$NMF on data sets, and the experimental results are shown in Figs. 1-2. For our proposed algorithms, the convergence condition is set empirically to 10^{-6}.

The clustering results determined by parameters α and β are in Fig. 1. It is observed from Fig. 1 that when the values of α and β are too large or too small, it is not conducive to the clustering effect of RFCM-$L_{2,1}$NMF. For data sets with less noise disturbance and simple data structure, such as: for datasets Iris, Wine and Yale, the values of parameters α and β are selected on the interval $\left[10^{-3}, 10^{-2}\right]$. For data sets with more noise and more complex data structure, such as: staimage and USPS , the values of parameters α and β are selected on the interval $[10^{-1}, 1]$, this is because when the data set contains more noise, the noise control item will play a greater role, and the corresponding control parameters will also become larger. The clustering results determined by parameters m and ξ are in Fig. 2, where α and β are set to optimal values. We can observe that the overall trend is relatively stable in Fig. 2, which means that when α and β are optimal, m and ξ have little influence on the effect of the experiment, the value ranges of parameters m and ξ are $[1.4, 1.8]$ and $\left[10^{-1}, 1\right]$ respectively.

TABLE 2 Experimental results on seven data sets for the different algorithms,and the best results are shown in boldface.

Data sets	Evaluate criteria	FCM	LFCM	MSFCM	$L_{2,1}$ NMF	RFCM-$L_{2,1}$NMF
Iris	**F***	0.8852	0.8753	0.8917	0.8389	**0.9654**
	ACC	0.8867	0.7989	0.8933	0.8420	**0.9557**
	NMI	0.7419	0.7513	0.7581	0.6912	**0.8642**
Wine	**F***	0.7099	0.6993	0.7149	0.7898	**0.9154**
	ACC	0.6966	0.7089	0.3898	0.8633	**0.9257**
	NMI	0.4213	0.4501	0.4288	0.5823	**0.6510**
satimage	**F***	0.5533	0.5533	0.6228	0.3977	**0.8324**
	ACC	0.2917	0.2917	0.3898	0.3923	**0.7143**
	NMI	0.4501	0.4501	0.4912	0.3213	**0.6510**
yale	**F***	0.1697	0.1681	0.1784	0.4385	**0.4821**
	ACC	0.1273	0.1247	0.1201	0.4060	**0.4903**
	NMI	0.1279	0.1379	0.1367	0.4449	**0.4408**
USPS	**F***	0.3971	0.4043	0.3946	0.1905	**0.5903**
	ACC	0.3585	0.4185	0.3606	0.2064	**0.5854**
	NMI	0.3021	0.2955	0.3095	0.1514	**0.4408**

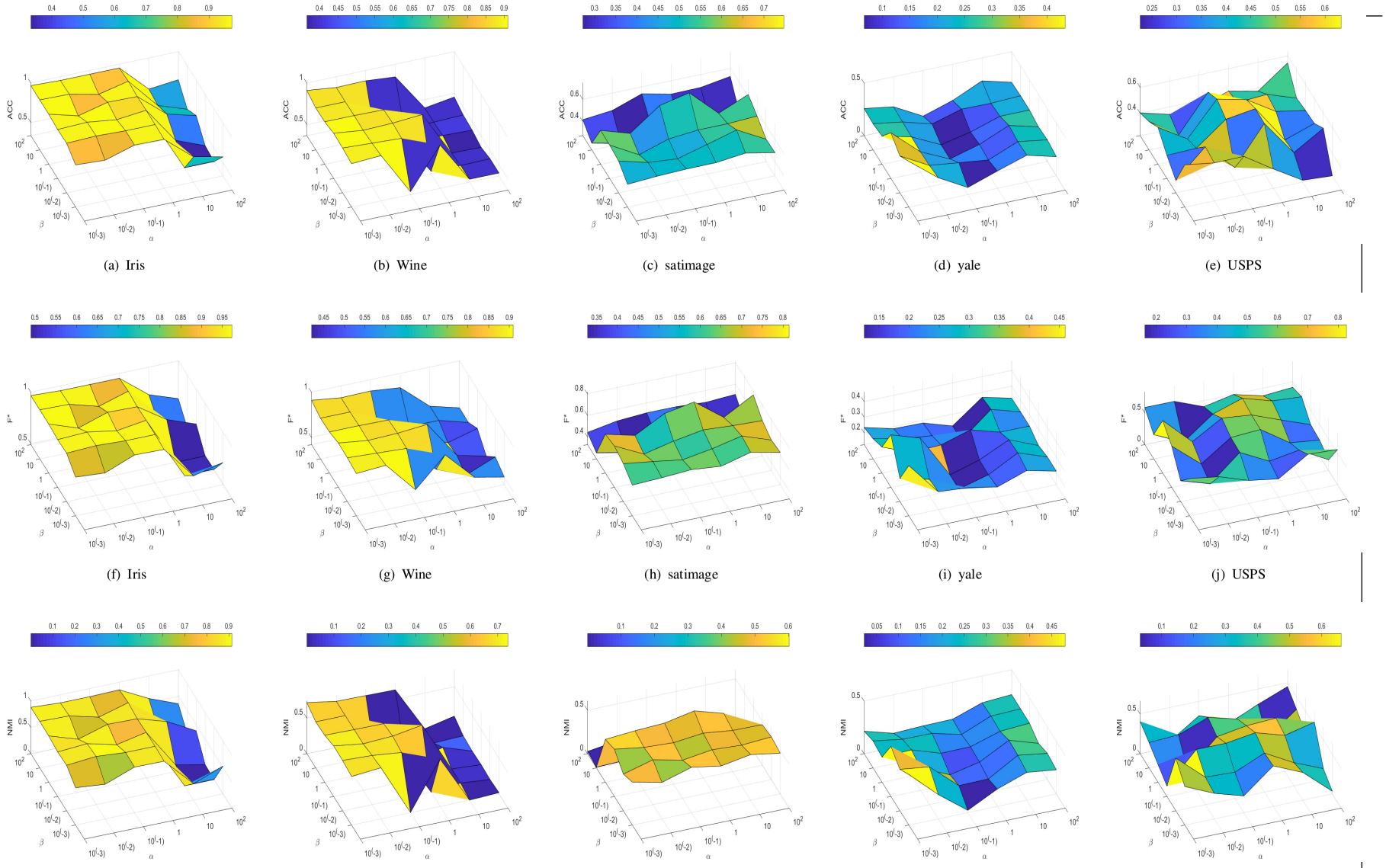
4.3 | Experimental analysis

As can be seen from Table 2, the experimental results of RFCM-$L_{2,1}$NMF in the five data sets have good clustering performance. The mean value of the performance are reported in Table 2 ,of which the best results are highlighted in boldface. From the experimental results, we have the following findings.

First, the results in the last column indicate that, under the above-mentioned settings, RFCM-$L_{2,1}$NMF outperforms the other algorithms on almost all data sets in terms of **F***, **ACC**, and **NMI**. Hence, RFCM-$L_{2,1}$NMF algorithm achieves higher clustering quality. Second, compared with FCM and MSFCM methods, the RFCM-$L_{2,1}$NMF algorithm demonstrates better performance in the **F***, **NMI**, and **ARI** measurement. This is because of the $L_{2,1}$NMF and noise constraint terms in our method, which can better remove noise in the data sets. Compared with $L_{2,1}$NMF, our method utilizes local constraints to capture more precise local geometry, resulting in better clustering results.For example, the accuracy of RFCM-$L_{2,1}$NMF is 11.3% higher than that of $L_{2,1}$NMF on Iris and the NMI of RFCM-$L_{2,1}$NMF is 15.98% higher. than that of MSFCM on satimage.

5 | CONCLUSIONS

Based on $L_{2,1}$NMF and FCM, Robust denoising FCM clustering via $L_{2,1}$ NMF and local constraint (RFCM-$L_{2,1}$NMF) is proposed in this paper. First, RFCM-$L_{2,1}$NMF combines $L_{2,1}$NMF and FCM, and makes up for the shortcomings of FCM by using the



(a) Iris (b) Wine (c) satimage (d) yale (e) USPS

(f) Iris (g) Wine (h) satimage (i) yale (j) USPS

(k) Iris (l) Wine (m) satimage (n) yale (o) USPS

FIGURE 1 Clustering performance of RFCM-$L_{2,1}$NMF on five data sets versus parameters α and β. (a)-(e)**ACC**. (f)-(j)$\mathbf{F}^*$. (k)-(o) **NMI**.

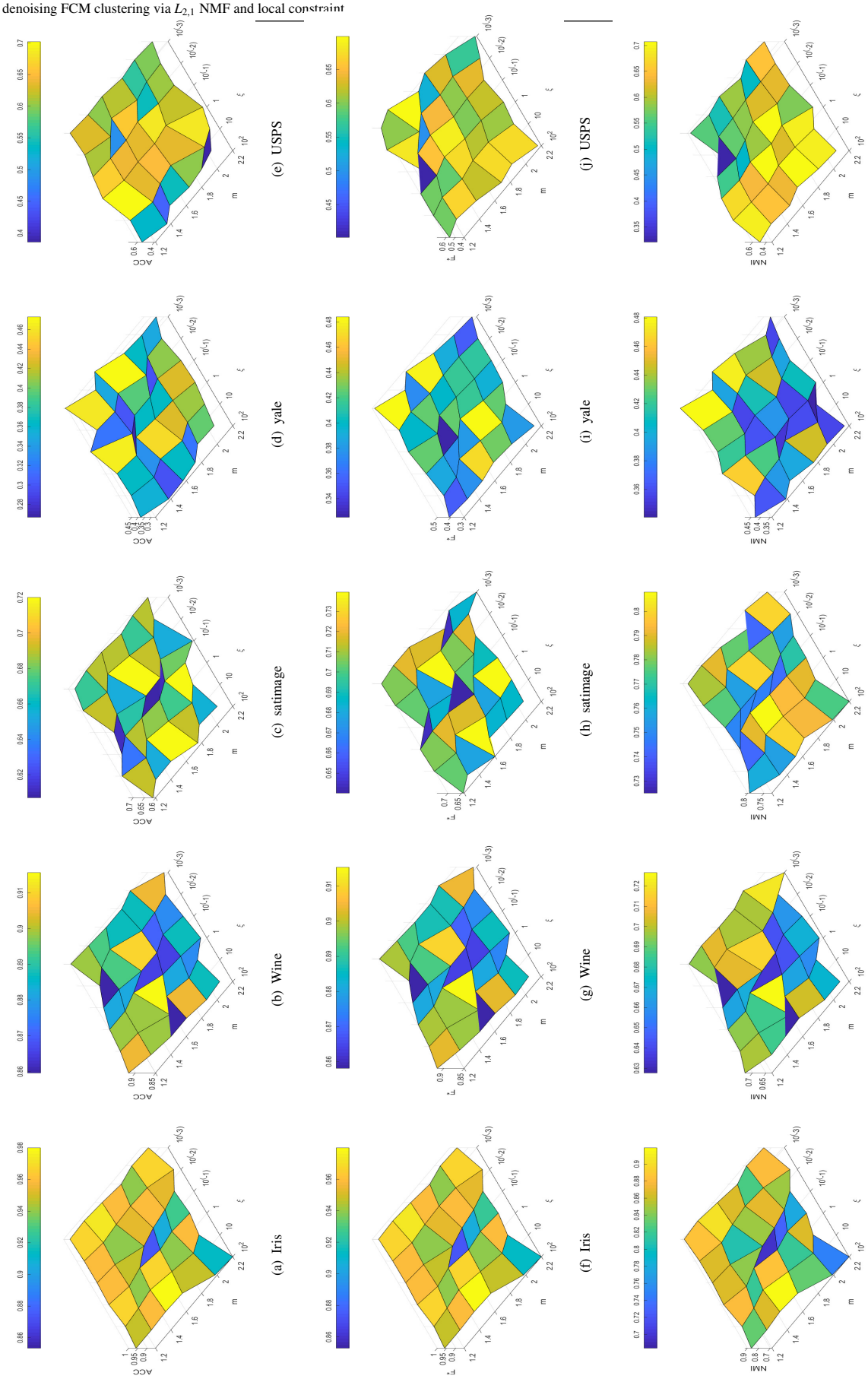


FIGURE 2 Clustering performance of RFCM-$L_{2,1}$NMF on five data sets versus parameters m and ξ. (a)-(e)**ACC**. (f)-(j)$\mathbf{F}^*$. (k)-(o) **NMI**.

advantages of column $L_{2,1}$NMF. Second, fuzzy clustering is used in the low-dimensional subspace of the original data, which avoids the FCM being affected to some extent by the initial value. At the same time, the $L_{2,1}$NMF with residual and noise constraint reduce the interference of noise. In addition, in order to strengthen the connection with the structural features of the data set, we construct a new local constraint term, in which an entropy-like divergence is used for the replacement of the commonly used Euclidean distance metric to further reduce the disturbance of noise, making the low-dimensional representation more efficient accurate.

AUTHOR CONTRIBUTIONS

This is an author contribution text. This is an author contribution text. This is an author contribution text. This is an author contribution text. This is an author contribution text.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China(11961010, 61967004).

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

REFERENCES

1. Kogan J, Nicholas C, Teboulle M, others . A Survey of Clustering Data Mining Techniques. *Springer.* 2006;pp:25–71.
2. Bezdek , James C. Pattern Recognition with Fuzzy Objective Function Algorithms. *Advanced Applications in Pattern Recognition.* 1981;22(1171):203–239.
3. Xu R, Wunsch D. Survey of clustering algorithms. *IEEE Transactions on Neural Networks.* 2005;16(3):645–678.
4. Bezdek JC, Ehrlich R, Full W. FCM: The fuzzy c -means clustering algorithm. *Computers and Geosciences.* 1984;10(2–3):191-203.
5. Dunn JC. A Graph Theoretic Analysis of Pattern Classification via Tamura's Fuzzy Relation. *IEEE Transactions on Systems Man and Cybernetics.* 1974;SMC-4(3):310-313.
6. Zadeh LA. Similarity relations and fuzzy orderings. *Information sciences.* 1971;3(2):177–200.
7. Wu Z, Leahy R. An optimal graph theoretic approach to data clustering: theory and its application to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 1993;15(11):1101-1113.
8. Hathaway RJ, Hu Y. Density-Weighted Fuzzy c-Means Clustering. *IEEE Transactions on Fuzzy Systems.* 2009;17(1):243-252.
9. Hungand M, Yang D. An intersection based ALE scheme (xALE) for cell centered hydrodynamics (CCH). In: IEEE. 2001:225–232
10. Gao Y, Wang D, Pan J, Wang Z, Chen B. A Novel Fuzzy c-Means Clustering Algorithm Using Adaptive Norm. *International Journal of Fuzzy Systems.* 2019;21(8):2632-2649.
11. Liu J, Fan J. A novel fuzzy c-means clustering algorithm based on local density. In: Springer. 2020:46–58.
12. Park D, Dagher I. Gradient based fuzzy c-means (GBFCM) algorithm. In: . 3. IEEE. 1994:1626-1631 vol.3
13. Havens TC, Bezdek JC, Leckie C, Hall LO, Palaniswami M. Fuzzy c-Means Algorithms for Very Large Data. *IEEE Transactions on Fuzzy Systems.* 2012;20(6):1130-1146.
14. Li C, Yu J. A Novel Fuzzy C-Means Clustering Algorithm. In: . 4062. Springer. 2006:510–515
15. Krinidis S, Chatzis V. A Robust Fuzzy Local Information C-Means Clustering Algorithm. *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society.* 2010;19(5):1328.
16. Zhou S, Li D, Zhang Z, Ping R. A New Membership Scaling Fuzzy C-Means Clustering Algorithm. *IEEE Transactions on Fuzzy Systems.* 2020;pp(99):1-1.
17. Wang C, Pedrycz W, Li ZW, Zhou MC. Residual-driven Fuzzy C-Means Clustering for Image Segmentation. *IEEE CAA journal of automatica sinica.* 2021;8(4):876 - 889.
18. Lee D. Learning parts of objects by non-negative matrix factorization. *Letter of Nature.* 1999;pp:788-791.
19. Sun L, Hongwei G, Kang W. Non-negative matrix factorization based modeling and training algorithm for multi-label learning. *Frontiers of Computer Science (print).* 2018;13(006):1243-1254.
20. Huck A, Guillaume M. Robust hyperspectral data unmixing with spatial and spectral regularized NMF. In: IEEE. 2010:1-4
21. Kong D, Ding C, Huang H. Robust nonnegative matrix factorization using l21-norm. In: ACM. 2011:673-682.
22. Hoyer P. Non-negative sparse coding. In: . pp. IEEE. 2002:557-565.
23. Deng C, He X, Han J. Graph Regularized Non-negative Matrix Factorization for Data Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2012;pp:1548-1560.
24. Ye W, Wang H, Shan Y, Li T, Yan Y. Nonnegative matrix factorization for clustering ensemble based on dark knowledge. *Knowledge-Based Systems.* 2018;163:624-631.
25. Tong M, Bai H, Yue X, Bu H. PTL-LTM model for complex action recognition using local-weighted NMF and deep dual-manifold regularized NMF with sparsity constraint. *Neural Computing and Applications.* 2020;32(17):13759-13781.
26. Tao X, Yu L, Wang X. One method based on non-negative matrix factorization and fuzzy C means for image clustering. *Information Technology and Network Security.* 2019;38(17):44-48.
27. Wu C, Cao Z. Noise distance driven fuzzy clustering based on adaptive weighted local information and entropy-like divergence kernel for robust image segmentation. *Digital Signal Processing.* 2021;111:102963.
28. Kriegel HP, Kröger P, Zimek A. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *Acm transactions on knowledge discovery from data (tkdd).* 2009;3(1):1-58.
29. Parker JK, Hall LO. Accelerating Fuzzy-C Means Using an Estimated Subsample Size. *IEEE Transactions on Fuzzy Systems.* 2014;17(5):1229-1244.

30. Bezdek J, Hathaway R, Sabin M, Tucker W. Convergence theory for fuzzy c-means: Counterexamples and repairs. *IEEE Transactions on Systems, Man, and Cybernetics.* 1987;17(5):873-877.
31. Pascual-Montano A, Carazo JM, Kochi K, Lehmann D, Pascual-Marqui RD. Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE transactions on pattern analysis and machine intelligence.* 2006;28(3):403–415.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.