Pattern Recognition 138 (2023) 109405

Contents lists available at ScienceDirect

Pattern Recognition

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

Robust weighted co-clustering with global and local discrimination

Zhoumin Lu[a,b,c], Shiping Wang[d,e], Genggeng Liu[d,e], Feiping Nie[a,b,c,*]

[a] School of Computer Science, Northwestern Polytechnical University, Xian 710072, PR China
[b] School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xian 710072, PR China
[c] Key Laboratory of Intelligent Interaction and Applications (Northwestern Polytechnical University), Ministry of Industry and Information Technology, Xian 710072, PR China
[d] College of Computer and Data Science, Fuzhou University, Fuzhou 350116, PR China
[e] Fujian Provincial Key Laboratory of Network Computing and Intelligent Information Processing, Fuzhou 350116, PR China

ARTICLE INFO

Article history:
Received 25 November 2021
Revised 29 October 2022
Accepted 5 February 2023
Available online 7 February 2023

Keywords:
Machine learning
Co-clustering
Nonnegative matrix factorization
Global discrimination
Local discrimination

ABSTRACT

In the past few decades, the clustering problem has made considerable progress, and co-clustering algorithms have attracted more attention. Compared with one-side clustering, co-clustering not only groups samples according to the distribution of features but also groups features according to the distribution of samples at the same time. This duality helps to explore the structural information of data, such as genes and texts. In this paper, a new co-clustering algorithm is proposed to simultaneously consider feature weights, data noise, local manifolds, and global scatter, named robust weighted co-clustering with global and local discrimination. Furthermore, an alternate update rule is put forward to optimize objective, theoretically proven to converge. Then, the algorithm's duality, robustness, and effectiveness have been verified on synthetic, corrupted, and real datasets, respectively. The runtime and parameter sensitivity of the algorithm are also analyzed. Finally, sufficient experiments clarify the competitiveness of our algorithm compared to other ones.

1. Introduction

The research on clustering has been enduring for a long time and is still receiving widespread attention today. Specifically, clustering is a process and method of dividing data according to a certain standard. It tends to group data with similar attributes into one category. For intricate data, if you want to reflect specific characteristics or rules, it is often necessary to resort to cluster analysis methods.

It is worth noting that common one-side clustering algorithms only take into account the distribution of samples on the features, such as k-means. However, current research indicates that co-clustering generally performs better than one-side clustering. In the process of co-clustering, samples are grouped according to their distribution on features, and features are also grouped according to their distribution on samples. This parallel approach potentially takes into account the internal connections and interactions between samples and features, and is more suitable for structured data. Therefore, it is used in text mining [1–3], recommendation systems [4–6], gene expression [7,8], collaborative filtering [9,10], image processing [11–13], social relationship detection [14,15], interactive network recognition [16,17], etc. The application field is quite successful.

In this paper, we propose a robust weighted co-clustering algorithm that simultaneously considers feature weights, data noise, local manifolds, and global scatter. Furthermore, an alternate update rule is put forward to optimize objective, theoretically proven to converge. Then, the algorithm's duality, robustness, and effectiveness have been verified on comprehensive experiments, showing the superior performance of our method. Overall, we summarize the main contributions as follows.

- Form a weighted 0-norm non-negative matrix factorization with global and local discriminant regularizers to simultaneously deal with multiple problems in co-clustering.
- Optimize the formed objective function for solving the proposed co-clustering problem by an alternate update rule, proven to converge in theory.
- Verify our algorithm's duality, robustness, effectiveness, and competitiveness over compared ones through comprehensive experiments.

[*] Corresponding author.
E-mail addresses: walker.zhoumin.lu@gmail.com (Z. Lu), feipingnie@gmail.com (F. Nie).

https://doi.org/10.1016/j.patcog.2023.109405
0031-3203/© 2023 Elsevier Ltd. All rights reserved.

2. Related work

In the past two decades, non-negative matrix factorization (NMF) [18,19] has received more and more attention. It provides a large number of techniques to describe and optimize machine learning problems, especially for constrained clustering tasks. It is worth mentioning that both k-means clustering and spectral clustering are closely related to NMF, and can be expressed as a constraint form of NMF [20,21]. The original NMF was designed to seek a good low-rank approximation, potentially realizing the role of dimensionality reduction. Compared with singular value decomposition and principal component analysis, its non-negative characteristics are often more suitable for real life. In addition, semi-nonnegative matrix factorization (SNMF) [22] and graph-regularized non-negative matrix factorization (GNMF) [23] are powerful extensions of this technology. SNMF relaxes the non-negativity of one of the components and treats it as a cluster center matrix, thus directly transforming the process of learning low-dimensional representation into a clustering process. GNMF maintains a local manifold in the process of learning low-dimensional representations, thereby improving the effectiveness of new features.

Co-clustering can generally be divided into two categories: NMF-based and bipartite-graph-based methods. In NMF-based methods, DRCC [24] and DNMTF [25] maintain the local manifold structure of samples and features, ONMTF [26], DLLC [27], and PNMF [28] introduce orthogonal constraints, FNMTF [29] focuses on fast algorithm solving, NMTFCoS [30] co-shrinks irrelevant features by encouraging co-sparseness of model parameters, while MultiCC [31] expects to maximize the difference between each division. In bipartite-graph-based methods, LDCC [32] takes into account the local discrimination, BKM [33] attempts to solve quickly, while SOBG [34] focuses on the search for an optimal structured bipartite graph. In addition, there are some co-clustering methods that do not fall into the above two categories. For example, information-theory-based ITCC [35] minimizes the mutual information loss between row-to-column and column-to-row clusters. Ensemble-learning-based CoCE [36] simultaneously measures feature-to-feature, object-to-object and feature-to-object relevance information. Swarm-intelligence-based co-ABC [37] develops an artificial bee colony algorithm with co-similarity metric and local search.

3. Proposed method

In this section, a robust weighted co-clustering with global and local discrimination is put forward and then solved by iterative update rules, guaranteed to converge in theory.

3.1. Problem formulation

Suppose a data matrix $X \in \mathbb{R}^{d\times n}$ is given, where d denotes the feature dimension and n represents the number of samples. If the feature space and the sample space are marked as $\mathcal{Y} = \{y_1,\dots,y_d\}$ and $\mathcal{X} = \{x_1,\dots,x_n\}$, then $X = (x_1,\dots,x_n) = (y_1,\dots,y_d)^T = Y^T$ holds. According to the literature [38], the data matrix X can be regarded as composed of two parts: structural information L and irrelevant noise E. Generally, the internal structure information is row-related or column-related, having low-rank characteristics, while noise is often sparse. Therefore, the following relationship holds.

$$\min \|E\|_0 + \lambda\, \mathrm{rank}(L) \qquad \text{s.t. } X = L + E \tag{1}$$

In general co-clustering, it is usually considered to decompose the data matrix X directly. If data noise is considered, the following relationship is more reasonable.

$$L = FSG^T \qquad \text{s.t. } F \geq 0, G \geq 0 \tag{2}$$

where S is a shared coefficient, F and G are indicators for features and samples, respectively. Substituting Eq. (2) into Eq. (1), then Eq. (3) can be approximated. Since rank(L) is not greater than a certain constant, it is omitted from the formula.

$$\min_{F,S,G} \left\|X - FSG^T\right\|_0 \qquad \text{s.t. } F \geq 0, G \geq 0 \tag{3}$$

3.2. Global discrimination

If the centering matrix $H_n = I - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ is given for sample space, the data can be rearranged in the form of $\hat{X} = XH_n$, and $\sum_{i=1}^n \hat{x}_i/n = 0$ holds. According to related documents [39,40], the inter-class scatter S_b^g and total scatter S_t^g of samples are defined as

$$S_b^g = \hat{X}GG^T\hat{X}^T \tag{4}$$

$$S_t^g = \hat{X}\hat{X}^T \tag{5}$$

Furthermore, we map the original data X into a high-dimensional space for linear separability. Then kernel trick is applied for a solution. Let $\phi(\cdot)$ be a mapping function and $\tilde{X} = \phi(X)H_n$, the inter-class scatter $\tilde{S}_b^g$ and total scatter $\tilde{S}_t^g$ of high-dimensional samples are defined as

$$\tilde{S}_b^g = \tilde{X}GG^T\tilde{X}^T \tag{6}$$

$$\tilde{S}_t^g = \tilde{X}\tilde{X}^T \tag{7}$$

Generally speaking, the sample spacing between clusters should be as large as possible, while the sample spacing within the cluster should be as small as possible. Under such a standard, maximizing the following goal (8) is reasonable, where μ is a positive parameter used to balance the two scatter matrices and make the matrix invertible. For subsequent joint optimization, Eq. (8) is equivalently transformed as

$$\max_G \mathrm{Tr}\left[\left(\tilde{S}_t^g + \mu I\right)^{-1}\tilde{S}_b^g\right] \tag{8}$$

$$\Leftrightarrow \max_G \mathrm{Tr}\left[G^T\tilde{X}^T\left(\tilde{X}\tilde{X}^T + \mu I\right)^{-1}\tilde{X}G\right] \tag{9}$$

$$\Leftrightarrow \max_G \mathrm{Tr}\left[G^T\left(\tilde{X}^T\tilde{X} + \mu I\right)^{-1}\tilde{X}^T\tilde{X}G\right] \tag{10}$$

$$\Leftrightarrow \max_G \mathrm{Tr}\left\{G^T\left[\left(\mu I + \tilde{X}^T\tilde{X}\right)^{-1}\left(\tilde{X}^T\tilde{X} + \mu I - \mu I\right)\right]G\right\} \tag{11}$$

$$\Leftrightarrow \max_G \mathrm{Tr}\left\{G^T\left[I - \left(I + \frac{1}{\mu}\tilde{X}^T\tilde{X}\right)^{-1}\right]G\right\} \tag{12}$$

$$\Leftrightarrow \min_G \mathrm{Tr}\left\{G^T\left[\left(I + \frac{1}{\mu}H_n^TK_GH_n\right)^{-1} - I\right]G\right\} \tag{13}$$

$$\Leftrightarrow \min_G \mathrm{Tr}\left(G^TR_GG\right) \tag{14}$$

where $R_G = \left(I + \frac{1}{\mu}H_n^TK_GH_n\right)^{-1} - I$ and $K_G = \phi(X)^T\phi(X) = K(X,X)$ is a kernel function.

Similarly, given centering matrix $H_d = I - \frac{1}{d}\mathbf{1}_d\mathbf{1}_d^T$ for feature space and let $\tilde{Y} = \phi(Y)H_d$, the inter-class scatter $\tilde{S}_b^f$ and total scatter $\tilde{S}_t^f$ of mapped features are defined as

$$\tilde{S}_b^f = \tilde{Y}FF^T\tilde{Y}^T \tag{15}$$

$$\tilde{S}_t^f = \tilde{Y}\tilde{Y}^T \tag{16}$$

It is desirable to achieve the following goal (17) for large inter-class scatter and small total scatter of mapped feature points, where $K_F = \phi(Y)^T\phi(Y) = K(Y,Y)$ is a kernel function and $R_F = \left(I + \frac{1}{\mu}H_d^T K_F H_d\right)^{-1} - I$.

$$\max_F \mathrm{Tr}\left[\left(\tilde{S}_t^f + \mu I\right)^{-1}\tilde{S}_b^f\right] \tag{17}$$

$$\Leftrightarrow \min_F \mathrm{Tr}\left\{F^T\left[\left(I + \frac{1}{\mu}H_d^T K_F H_d\right)^{-1} - I\right]F\right\} \tag{18}$$

$$\Leftrightarrow \min_F \mathrm{Tr}\left(F^T R_F F\right) \tag{19}$$

3.3. Local discrimination

The local data structure should also be considered in the clustering process, even more important than the global structure. In fact, the graph is a natural local structure, which measures the adjacency relationship between two objects. Therefore, graph structures are widely used in feature selection, dimensionality reduction and image retrieval, all of which have apparent performance improvements. However, the shortcomings of graph structure are also evident. It only describes the binary relation between objects, while objects in reality usually have an n-ary association, such as literature citation. Unlike the traditional graph structure, the hyperedge in hypergraph can contain multiple objects, describing the n-ary relation among objects. It is a high-dimensional local structure representation method.

General data does not have an explicit hypergraph structure, so this subsection adopts an artificial construction method. First, each sample is regarded as the center to construct a hyperedge. Furthermore, the samples closest to the center are selected as hyperedge's members. We utilize the similarity between samples and hyperedge's centers to estimate the probability that each sample belongs to every hyperedge instead of the binary membership in the standard hypergraph. That is, the degree of membership relaxes to [0,1] interval as

$$h(x_i, e_j) = \begin{cases} \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right) & \text{if } x_i \in e_j \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

where $h(x_i, e_j)$ represents the membership degree of sample x_i concerning hyperedge e_j. In addition, the weight of the hyperedge is estimated by the compactness of the samples belonging to the hyperedge. The compactness is the mean value of the similarity of all samples within the hyperedge. The smaller the value, the more compact the distribution of these samples, and the stronger the existence of the hyperedge, as below:

$$w(e) = \frac{1}{|e|^2}\sum_{x_i, x_j \in e}\exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right) \tag{21}$$

So far, a complete hypergraph has been constructed. According to its definition, the degree of a vertex $d(x)$ is the sum of the weights of its associated hyperedges, and the degree of a hyperedge $\delta(e)$ is the sum of the membership degrees of its associated vertices, as follows:

$$d(x) = \sum_i w(e_i)h(x, e_i) \tag{22}$$

$$\delta(e) = \sum_i h(x_i, e) \tag{23}$$

Obviously, the higher the correlation between samples, the more likely they belong to the same category. On the contrary, inter-cluster samples should avoid associations, so there are the following:

$$\sum_{e\in E_g}\sum_{x_i, x_j \in V_g}\frac{w(e)h(x_i, e)h(x_j, e)}{\delta(e)}\left\|g_i - g_j\right\|_2^2 \tag{24}$$

$$= \mathrm{Tr}\left(G^T\left(D_{vG} - H_G W_G D_{eG}^{-1} H_G^T\right)G\right) \tag{25}$$

$$= \mathrm{Tr}\left(G^T L_G G\right) \tag{26}$$

where $L_G = \left(D_{vG} - H_G W_G D_{eG}^{-1} H_G^T\right)$, H_G is the membership degree matrix, W_G, D_{vG}, and D_{eG} are diagonal matrices composed of the hyperedge weight, the degree of the vertex, and the degree of the hyperedge, respectively. Similar to the case of samples, features also have such a local structure, and the following can be known:

$$\sum_{e\in E_f}\sum_{y_i, y_j, \in V_f}\frac{w(e)h(y_i, e)h(y_j, e)}{\delta(e)}\left\|f_i - f_j\right\|_2^2 \tag{27}$$

$$= \mathrm{Tr}\left(F^T\left(D_{vF} - H_F W_F D_{eF}^{-1} H_F^T\right)F\right) \tag{28}$$

$$= \mathrm{Tr}\left(F^T L_F F\right) \tag{29}$$

where $L_F = \left(D_{vF} - H_F W_F D_{eF}^{-1} H_F^T\right)$, H_F is the membership degree matrix, W_F, D_{vF}, and D_{eF} are diagonal matrices composed of the hyperedge weight, the degree of the vertex, and the degree of the hyperedge, respectively.

3.4. Objective function

As the central part of our objective function, Eq. (3) is an essential but difficult-to-optimize problem, so it is relaxed into the following form:

$$\begin{aligned} &\min\nolimits_{F,S,G,E}\left\|X - FSG^T - E\right\|_F^2 + \gamma\|E\|_0 \\ &\text{s.t. } F \ge 0, G \ge 0 \end{aligned} \tag{30}$$

The E here can be understood as directly modeling the noise. In addition, the importance of different features is often different, so the weights are added as below for self-learning during the optimization process.

$$\begin{aligned} &\min\nolimits_{F,S,G,E,W}\left\|W\left(X - FSG^T - E\right)\right\|_F^2 + \gamma\|E\|_0 \\ &\text{s.t. } \textstyle\sum_{m=1}^M w_m = d, W \ge 0, F \ge 0, G \ge 0 \end{aligned} \tag{31}$$

Combined with the global and local discrimination, the final objective function can be obtained as follows:

$$\begin{aligned} &\min\nolimits_{F,S,G,E,W}\left\|W\left(X - FSG^T - E\right)\right\|_F^2 \\ &\qquad + \mathrm{Tr}\left(F^T Q_F F\right) + \mathrm{Tr}\left(G^T Q_G G\right) + \gamma\|E\|_0 \\ &\text{s.t.} \quad \textstyle\sum_{i=1}^d w_i = d, W \ge 0, F \ge 0, G \ge 0 \end{aligned} \tag{32}$$

where W indicates feature weight whose non-diagonal elements are all 0, w_i is the ith diagonal element of W, γ is an adjustment parameter, $Q_F = \alpha_1 R_F + \beta_1 L_F$ and $Q_G = \alpha_2 R_G + \beta_2 L_G$.

3.5. Optimization algorithm

This subsection will solve the objective function by iterative optimization of S, F, G, W and E. Moreover, the label assignment method and algorithm summary will be given.

3.5.1. Update coefficient S

Denote $A = X - E$ and $D = W^T W$, the objective function for S can be rewritten as

$$\begin{aligned} J(S) &= \left\|W\left(A - FSG^T\right)\right\|_F^2 \\ &= \mathrm{Tr}\left[\left(A - FSG^T\right)^T D\left(A - FSG^T\right)\right] \\ &= \mathrm{Tr}\left(A^T DA\right) - 2\,\mathrm{Tr}\left(A^T\ \mathrm{DFSG}^T\right) + \mathrm{Tr}\left(GS^T F^T DFSG^T\right) \end{aligned} \tag{33}$$

By taking derivatives of $J(S)$ on S, we have

$$\frac{\partial J(S)}{\partial S} = -2F^T DAG + 2F^T DFSG^T G \tag{34}$$

Ulteriorly, setting $\frac{\partial J(S)}{\partial S} = 0$ results in

$$S = \left(F^T DF\right)^{-1}\left(F^T DAG\right)\left(G^T G\right)^{-1} \tag{35}$$

3.5.2. Update indicators F and G

First, rewrite the objective function of F into the following form

$$J(F) = \frac{1}{2}\left\|W\left(A - FSG^T\right)\right\|_F^2 + \frac{1}{2}\,\mathrm{Tr}\left(F^T Q_F F\right) \qquad \text{s.t. } F \geq 0 \tag{36}$$

For solving the constraints, we introduce the Lagrangian multiplier Φ and construct the Lagrangian function on F.

$$\mathcal{L}(F) = \frac{1}{2}\left\|W\left(A - FSG^T\right)\right\|_F^2 + \frac{1}{2}\,\mathrm{Tr}\left(F^T Q_F F\right) - \mathrm{Tr}(\Phi F) \tag{37}$$

By taking derivatives of $\mathcal{L}(F)$ on F, we have

$$\frac{\partial \mathcal{L}(F)}{\partial F} = -DAGS^T + DFSG^T GS^T + Q_F F - \Phi \tag{38}$$

Furthermore, denote $M = DAGS^T$ and $N = SG^T GS^T$, setting $\frac{\partial \mathcal{L}(F)}{\partial F} = 0$ results in

$$\Phi = -M + DFN + Q_F F \tag{39}$$

Combined with the KarushKuhnTucker conditions $\Phi_{ij} F_{ij} = 0$, we have

$$(-M + DFN + Q_F F)_{ij} F_{ij} = 0 \tag{40}$$

To ensure that each term in the above equation is non-negative, we introduce $M = M^+ - M^-$ and $N = N^+ - N^-$, then Eq. (40) can be rewritten into the following form.

$$\left(-M^+ + M^- + DFN^+ - DFN^- + Q_F^+ F - Q_F^- F\right)_{ij} F_{ij} = 0 \tag{41}$$

where $O^+ = \frac{|O|+O}{2}$ and $O^- = \frac{|O|-O}{2}$ for any matrix O. At this time, Eq. (41) satisfies the optimization framework of non-negative matrix factorization, and the following multiplicative update rule can be derived.

$$F_{ij} \leftarrow F_{ij}\left[\frac{\left(M^+ + DFN^- + Q_F^- F\right)_{ij}}{\left(M^- + DFN^+ + Q_F^+ F\right)_{ij}}\right]^{\frac{1}{2}} \tag{42}$$

Similar to the solution of F, the objective function on G can be rewritten as

$$J(G) = \frac{1}{2}\left\|W\left(A - FSG^T\right)\right\|_F^2 + \frac{1}{2}\,\mathrm{Tr}\left(G^T Q_G G\right) \qquad \text{s.t. } G \geq 0 \tag{43}$$

By introducting Lagrangian multiplier Ψ, the Lagrangian function on G can be constructed as

$$\mathcal{L}(G) = \frac{1}{2}\left\|W\left(A - FSG^T\right)\right\|_F^2 + \frac{1}{2}\,\mathrm{Tr}\left(G^T Q_G G\right) - \mathrm{Tr}(\Psi G) \tag{44}$$

By taking derivatives of $\mathcal{L}(G)$ on G, we have

$$\frac{\partial \mathcal{L}(G)}{\partial G} = -A^T DFS + GS^T F^T DFS + Q_G G - \Psi \tag{45}$$

Furthermore, denote $U = A^T DFS$ and $V = S^T F^T DFS$, setting $\frac{\partial \mathcal{L}(G)}{\partial G} = 0$ results in

$$\Psi = -U + GV + Q_G G \tag{46}$$

Combined with the KarushKuhnTucker conditions $\Psi_{ij} G_{ij} = 0$, we have

$$(-U + GV + Q_G G)_{ij} G_{ij} = 0 \tag{47}$$

To ensure that each term in the above equation is non-negative, we introduce $U = U^+ - U^-$ and $V = V^+ - V^-$, then Eq. (47) can be rewritten into the following form.

$$\left(-U^+ + U^- + GV^+ - GV^- + Q_G^+ G - Q_G^- G\right)_{ij} G_{ij} = 0 \tag{48}$$

At this time, Eq. (48) satisfies the optimization framework of non-negative matrix factorization, and the following multiplicative update rule can be derived.

$$G_{ij} \leftarrow G_{ij}\left[\frac{\left(U^+ + GV^- + Q_G^- G\right)_{ij}}{\left(U^- + GV^+ + Q_G^+ G\right)_{ij}}\right]^{\frac{1}{2}} \tag{49}$$

3.5.3. Update weight W

Remove irrelevant terms, and rewrite the objective function on W into the following form.

$$\begin{aligned} &J(W) = \mathrm{Tr}\left(W^T WZZ^T\right) = \textstyle\sum_{i=1}^d w_i^2 \sum_{j=1}^n z_{ij}^2 = \sum_{i=1}^d w_i^2 p_i \\ &\text{s.t.} \quad \textstyle\sum_{i=1}^d w_i = d, W \geq 0 \end{aligned} \tag{50}$$

where $Z = A - FSG^T$, $p_i = \sum_{j=1}^n z_{ij}^2$ and z_{ij} is an element of Z. According to the Cauchy-Schwarz inequality:

$$\sum_{k=1}^n a_k^2 \sum_{k=1}^n b_k^2 \geq \left(\sum_{k=1}^n a_k b_k\right)^2 \tag{51}$$

the following unequal relation is constructed:

$$d^2 = \left(\sum_{i=1}^d w_i\right)^2 = \left(\sum_{i=1}^d w_i\sqrt{p_i}\frac{1}{\sqrt{p_i}}\right)^2 \leq \left(\sum_{i=1}^d w_i^2 p_i\right)\left(\sum_{i=1}^d \frac{1}{p_i}\right) \tag{52}$$

After transposition of terms, we know

$$J(W) = \sum_{i=1}^d w_i^2 p_i \geq \frac{d^2}{\sum_{i=1}^d \frac{1}{p_i}} \tag{53}$$

If and only if Eq. (54) is established, the function takes the minimum value, where C is a constant.

$$w_i\sqrt{p_i} = C\frac{1}{\sqrt{p_i}} \Rightarrow w_i = C\frac{1}{p_i} \tag{54}$$

Combining the constraints on W, this constant can be calculated as below.

$$C\sum_{i=1}^d \frac{1}{p_i} = \sum_{i=1}^d w_i = d \Rightarrow C = \frac{d}{\sum_{i=1}^d \frac{1}{p_i}} \tag{55}$$

Finally, by substituting Eq. (55) into Eq. (54), the solution of W is obtained.

$$w_i = \frac{d/p_i}{\sum_{i=1}^d 1/p_i} \tag{56}$$

3.5.4. Update noise E

Denote $B = W(X - FSG^T)$ and remove irrelevant terms. The objective function on E is rewritten as follows.

$$\begin{aligned} J(E) &= \|B - WE\|_F^2 + \gamma\|E\|_0 \\ &= \textstyle\sum_{ij}\left[\left(b_{ij} - w_i e_{ij}\right)^2 + \gamma\left|e_{ij}\right|_0\right] \\ &= \textstyle\sum_{ij} f\left(e_{ij}\right) \end{aligned} \tag{57}$$

where b_{ij} and e_{ij} are the elements of B and E respectively, and w_i is the diagonal element of W. Further analysis of $f(e_{ij})$ shows that when $e_{ij} = 0$, $f(e_{ij}) = b_{ij}^2$, otherwise $f(e_{ij}) \geq \gamma$.

$$f\left(e_{ij}\right) = \begin{cases} b_{ij}^2, & \text{if } e_{ij} = 0 \\ \left(b_{ij} - w_i e_{ij}\right)^2 + \gamma\left|e_{ij}\right|_0, & \text{if } e_{ij} \neq 0 \end{cases} \tag{58}$$

Therefore, if $b_{ij}^2 = \gamma$, the minimum reaches at $e_{ij} = 0$. Otherwise, take the minimum when $b_{ij} - w_i e_{ij} = 0$.

$$E_{ij} = \begin{cases} \frac{b_{ij}}{w_i}, & \text{if } \left|b_{ij}\right| \geq \sqrt{\gamma} \\ 0, & \text{otherwise} \end{cases} \tag{59}$$

3.5.5. Label assignment

Through the iterative update, feature indicator F and sample indicator G gradually converge to $\widetilde{F}$ and $\widetilde{G}$. Further, the label assignment method for feature point is as follows.

$$l(y_i) = \arg\max_j \widetilde{F}_{ij} \tag{60}$$

Similarly, the label assignment method for sample point is as follows.

$$l(x_i) = \arg\max_j \widetilde{G}_{ij} \tag{61}$$

Conclusively, we employ Algorithm 1 to summarize the complete optimization process from the above analysis.

Algorithm 1 Robust weighted co-clustering with global and local discrimination (RWGLCC).

Input: Data matrix $X \in \mathbb{R}^{d\times n}$, feature cluster numbers c_1, sample cluster numbers c_2, parameters α_1, α_2, β_1, β_2 and γ.
Output: Feature labels $\{l(y_i)\}_{i=1}^d$ and sample labels $\{l(x_i)\}_{i=1}^n$.

```
Initialize $F$, $G$, $W$ and $E$;
Introduce global discrimination regularizers $R_F$ and $R_G$;
Introduce local discrimination regularizers $L_F$ and $L_G$;
while non-convergence do
    Compute $S = \left(F^T DF\right)^{-1}\left(F^T DAG\right)\left(G^T G\right)^{-1}$;
    Update $F_{ij} \leftarrow F_{ij}\left[\frac{(M^+ + DFN^- + Q_F^- F)_{ij}}{(M^- + DFN^+ + Q_F^+ F)_{ij}}\right]^{\frac{1}{2}}$;
    Update $G_{ij} \leftarrow G_{ij}\left[\frac{(U^+ + GV^- + Q_G^- G)_{ij}}{(U^- + GV^+ + Q_G^+ G)_{ij}}\right]^{\frac{1}{2}}$;
    Compute $w_i = \frac{d/p_i}{\sum_{i=1}^d 1/p_i}$;
    Compute $E_{ij} = \begin{cases} \frac{b_{ij}}{w_i}, & \text{if } \left|b_{ij}\right| \geq \sqrt{\gamma} \\ 0, & \text{otherwise} \end{cases}$;
end while
Assign feature labels $l(y_i)$ and sample labels $l(x_i)$.
```

3.6. Convergence analysis

Formula (32) is a multivariate problem whose global optimal solution is challenging to obtain. Therefore, we use iterative optimization to find its local optimal solution. When other variables are fixed, the optimal solutions for S, W and E can be obtained directly, but not for F and G. Hence, it is necessary to prove that updating rules (42) and (49) are monotonically non-increasing. For the convenience of the following statements, first give the relevant definition and lemmas.

Definition 1. [19] If $Z(h, h') \geq F(h)$ and $Z(h, h) = F(h)$, then $Z(h, h')$ is an auxiliary function of $F(h)$.

Lemma 1. *[19] When $Z(h, h')$ is an auxiliary function of $F(h)$, $F(h)$ does not increase under the update rule $h^{(t+1)} = \arg\min_h Z(h, h^{(t)})$.*

Lemma 2. *[22] $\forall A \in \mathbb{R}_+^{n\times n}, B \in \mathbb{R}_+^{k\times k}, S \in \mathbb{R}_+^{n\times k}, S' \in \mathbb{R}_+^{n\times k}$, and A, B are symmetric, the following inequality holds.*

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{(AS'B)_{ij}S_{ij}^2}{S'_{ij}} \geq \operatorname{Tr}(S^T ASB) \tag{62}$$

Theorem 1. *Denote*

$$J(F) = -\operatorname{Tr}\left(F^T M\right) + \tfrac{1}{2}\operatorname{Tr}\left(F^T DFN\right) + \tfrac{1}{2}\operatorname{Tr}\left(F^T Q_F F\right) \tag{63}$$

The auxiliary function of $J(F)$ can be shown as

$$\begin{aligned} Z\left(F, F'\right) \quad &= -\textstyle\sum_{i=1}^d \sum_{j=1}^{c_1} M_{ij}^+ F'_{ij}\left(1 + \log\frac{F_{ij}}{F'_{ij}}\right) + \sum_{i=1}^d \sum_{j=1}^{c_i} M_{ij}^- \frac{F_{ij}^2 + F_{ij}'^2}{2F'_{ij}} \\ &+ \tfrac{1}{2}\textstyle\sum_{i=1}^d \sum_{j=1}^{c_1} \frac{\left(DF'N^+\right)_{ij} F_{ij}^2}{F'_{ij}} + \frac{1}{2}\sum_{i=1}^d \sum_{j=1}^{c_1} \frac{\left(Q_F^+ F'\right)_{ij} F_{ij}^2}{F'_{ij}} \\ &- \tfrac{1}{2}\textstyle\sum_{i=1}^d \sum_{j=1}^{c_1} \sum_{k=1}^{c_1} N_{jk}^- D_{ii} F'_{ij} F'_{ik}\left(1 + \log\frac{F_{ij}F_{ik}}{F'_{ij}F'_{ik}}\right) \\ &- \tfrac{1}{2}\textstyle\sum_{i=1}^d \sum_{j=1}^{c_1} \sum_{k=1}^{c_1} \left(Q_F^-\right)_{jk} F'_{ji} F'_{ki}\left(1 + \log\frac{F_{ji}F_{ki}}{F'_{ji}F'_{ki}}\right) \end{aligned} \tag{64}$$

Obviously, $Z\left(F, F'\right)$ is convex in F, whose global minima is

$$F_{ij} = \arg\min_{F_i} Z\left(F, F'\right) = F'_{ij}\left[\frac{\left(M^+ + DFN^- + Q_F^- F\right)_{ij}}{\left(M^- + DFN^+ + Q_F^+ F\right)_{ij}}\right]^{\frac{1}{2}} \tag{65}$$

Proof. See Appendix A. □

Theorem 2. *Denote*

$$J(G) = -\operatorname{Tr}\left(G^T U\right) + \frac{1}{2}\operatorname{Tr}\left(G^T GV\right) + \frac{1}{2}\operatorname{Tr}\left(G^T Q_G G\right) \tag{66}$$

The auxiliary function of $J(G)$ can be shown as

$$\begin{aligned} Z\left(G, G'\right) &= -\textstyle\sum_{i=1}^n \sum_{j=1}^{c_2} U_{ij}^+ G'_{ij}\left(1 + \log\frac{G_{ij}}{G'_{ij}}\right) + \sum_{i=1}^n \sum_{j=1}^{c_2} U_{ij}^- \frac{G_{ij}^2 + G_{ij}'^2}{2G'_{ij}} \\ &+ \tfrac{1}{2}\textstyle\sum_{i=1}^n \sum_{j=1}^{c_2} \frac{\left(G'V^+\right)_{ij} G_{ij}^2}{G'_{ij}} + \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^{c_2} \frac{\left(Q_G^+ G'\right)_{ij} G_{ij}^2}{G'_{ij}} \\ &- \tfrac{1}{2}\textstyle\sum_{i=1}^n \sum_{j=1}^{c_2} \sum_{k=1}^{c_2} V_{jk}^- G'_{ij} G'_{ik}\left(1 + \log\frac{G_{ij}G_{ik}}{G'_{ij}G'_{ik}}\right) \\ &- \tfrac{1}{2}\textstyle\sum_{i=1}^n \sum_{j=1}^{c_2} \sum_{k=1}^{c_2} \left(Q_G^-\right)_{jk} G'_{ji} G'_{ki}\left(1 + \log\frac{G_{ji}G_{ki}}{G'_{ji}G'_{ki}}\right) \end{aligned} \tag{67}$$

Obviously, $Z\left(G, G'\right)$ is convex in G, whose global minima is

$$G_{ij} = \arg\min_{G_{ij}} Z\left(G, G'\right) = G'_{ij}\left[\frac{\left(U^+ + GV^- + Q_G^- G\right)_{ij}}{\left(U^- + GV^+ + Q_G^+ G\right)_{ij}}\right]^{\frac{1}{2}} \tag{68}$$

Proof. Evidenced with the proof of Theorem 1. □

4. Experiments

This section verifies our algorithm on synthetic, corrupted and real datasets for duality, robustness and effectiveness tests. In addition, parameter sensitivity analysis and hypergraph construction discussion are also carried out.

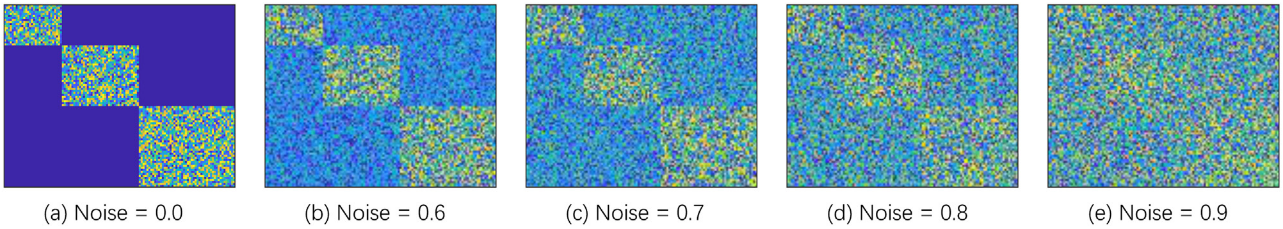


Fig. 1. A synthetic dataset composed of data matrices at different noise rates for duality test.

Table 1
The statistics on multiple real datasets.

Dataset	Type	#Instances	#Features	#Classes
COIL20	Object	1440	1024	20
USPS	Digit	9298	256	10
JAFFE	Face	213	4096	10
LUNG	Gene	203	3312	5
ISOLET	Speech	1560	617	26
BASEHOCK	Text	1993	4862	2

4.1. Dataset

For different purposes, synthetic data, corrupted data, and real data are used, respectively. The synthetic data is constructed as a two-dimensional matrix. As shown in Fig. 1, the rows and columns come from three clusters, and each block sub-matrix contains a row cluster and a column cluster. Besides, the elements in blocks are all sampled from the uniform distribution $U(0,1)$, and the area outside blocks is added with noise. The noise is also sampled from $U(0,1)$ but multiplied by a ratio. The corrupted data is extracted from the AT&T face dataset, and a random square area of each image is filled with black, referring to Fig. 2. The real data consists of six publicly available datasets from multiple sources, including object images, digital images, face images, gene expression, speech recognition and text documents. For instance, ISOLET is a speech recognition dataset, containing speech records where 30 subjects uttered 26 letters twice. Table 1 summarizes the details.

4.2. Setting

In order to verify that co-clustering can perform clustering from two directions simultaneously and achieve better performance, the duality test compares k-means [41], SNMF [22], NMTF [29] with RWGLCC on synthetic data by accuracy (ACC). Besides, k-means and SNMF need to cluster rows and columns independently, while NMTF and RWGLCC can group rows and columns simultaneously.

For the purpose of showing our algorithm robustness, let $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = 0$ and fix $W = I$, our objective function degenerates into the 0-norm non-negative matrix tri-factorization problem. In contrast, the original non-negative matrix tri-factorization problem employs F-norm. Further, they are compared on the corrupted dataset.

Considering the experimental comprehensiveness, we select some related one-side clustering and co-clustering methods for comparison, including SNMF [22], GNMF [23], FNMTF [29], DNMTF [25], BKM [33], SOBG [34], NMTFCoS [30] and MultiCC [31]. In detail, SNMF, FNMTF and BKM adopt the default settings. The regularization parameters of GNMF, DNMTF, SOBG, NMTFCoS, MultiCC and RWGLCC all use grid search by [0.001, 0.01, 0.1, 1, 10, 100, 1000].

In addition, we adopt k-means to initialize all NMF-based methods. Since the feature cluster counts are usually unknown, it is set to be consistent with the sample cluster counts. Graph construction follows the nearest neighbor algorithm. The nearest neighbor counts in feature space and sample space are set to 15. Taking into account the initialization sensitivity, we repeat all experiments 30 times, and record their average and standard deviation for comparison by accuracy, normalized mutual information (NMI) and adjusted rand index (ARI).

4.3. Result

As seen from Table 2, the clustering results of NMTF and RWGLCC in both directions are better than k-means and SNMF. This shows that NMTF and RWGLCC can indeed cluster at the same time and obtain better results. Besides, the performance of RWGLCC is better than that of NMTF, which indicates that our improvement is effective. In Fig. 2, F-norm NMTF has misclassified the face, while 0-norm NMTF is correct. This shows that 0-norm NMTF is more robust than F-norm NMTF, which in turn shows the robustness of our algorithm.

Relying on t-SNE [42], the clustering results on LUNG and USPS are visualized in Fig. 3. The clusters in LUNG have unbalanced sample numbers and obvious overlap. USPS has more samples, and some samples in clusters are far away. This makes most algorithm performance unsatisfactory. In contrast, the distribution of RWGLCC is closer to the distribution of GT, implying its better performance.

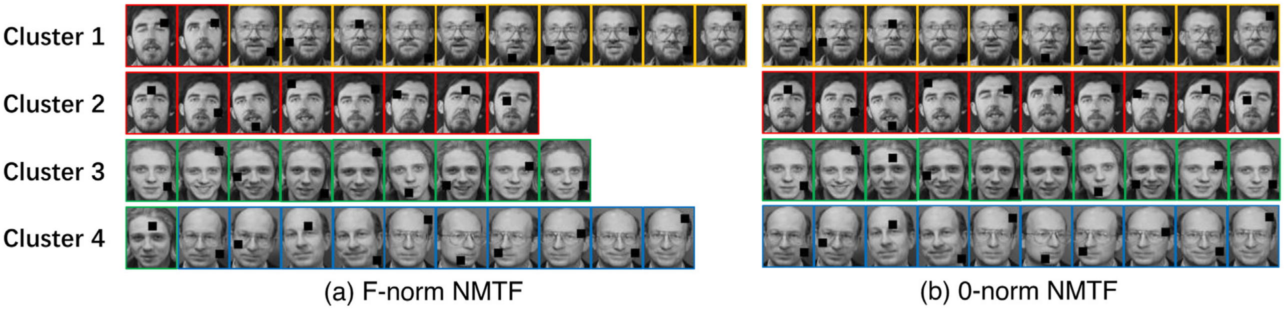


Fig. 2. Clustering results of F-norm and 0-norm NMTF on the corrupted dataset for robustness test.

Table 2
Duality test on the synthetic dataset by accurary (ACC%).

Direction	Method	Noise = 0.6	Noise = 0.7	Noise = 0.8	Noise = 0.9
Row	*k*-means	93.44	81.89	48.56	42.78
	SNMF	93.78	81.78	49.67	43.22
	NMTF	100.00	86.67	52.22	43.33
	RWGLCC	100.00	97.78	65.56	52.22
Column	*k*-means	91.58	80.50	47.33	42.58
	SNMF	94.58	81.50	50.25	43.92
	NMTF	99.17	87.50	51.67	44.17
	RWGLCC	100.00	93.33	65.83	53.33

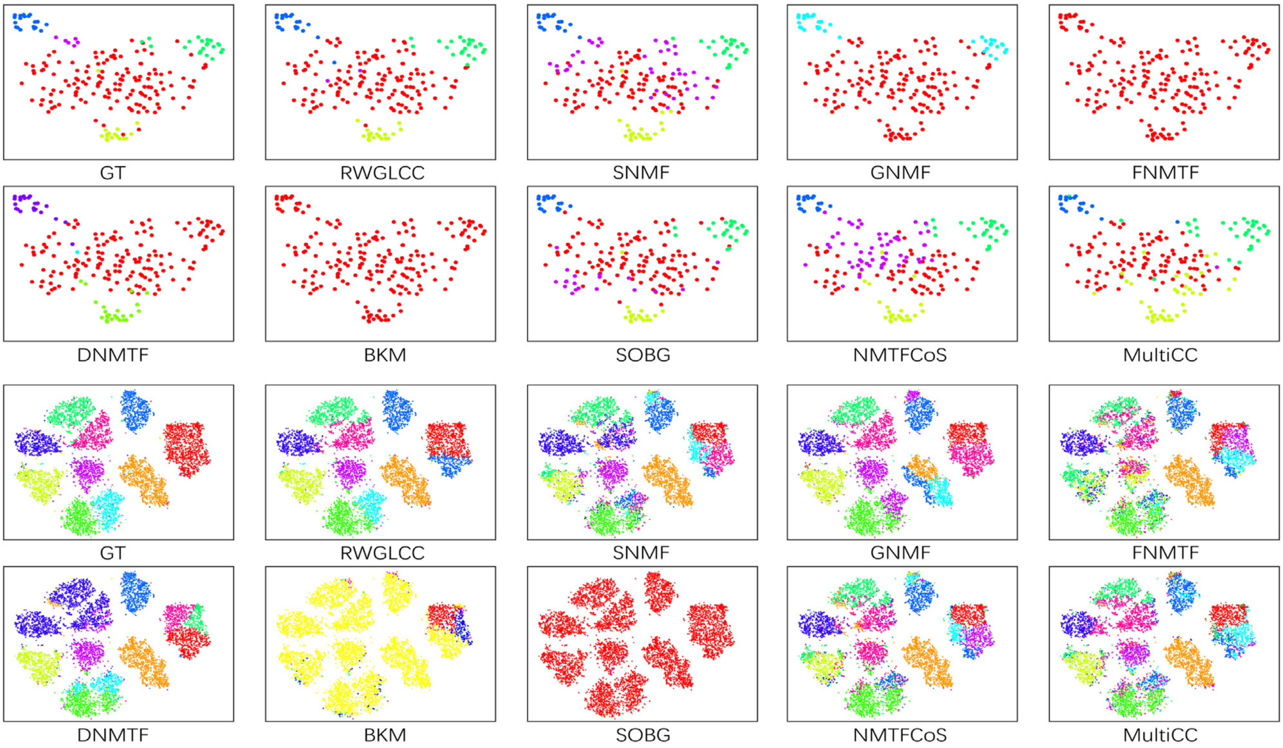


Fig. 3. Visualization for algorithm comparison on LUNG and USPS datasets.

Table 3
Effectiveness test on multiple real datasets by accuracy (ACC% ± std%).

Dataset	COIL20	USPS	JAFFE	LUNG	ISOLET	BASEHOCK
RWGLCC	**74.4 ± 3.1**	**80.9 ± 8.7**	**97.6 ± 0.8**	**93.8 ± 1.1**	**65.2 ± 2.0**	**91.5 ± 0.4**
SNMF	60.9 ± 4.2	64.3 ± 2.3	82.0 ± 8.9	76.8 ± 5.9	56.5 ± 3.1	64.9 ± 0.2
GNMF	65.0 ± 3.5	68.4 ± 9.0	86.3 ± 6.4	75.8 ± 7.2	57.0 ± 2.9	54.1 ± 7.6
FNMTF	61.8 ± 3.6	56.6 ± 2.2	77.2 ± 6.3	67.1 ± 4.4	54.0 ± 3.3	53.7 ± 0.0
DNMTF	65.4 ± 6.6	67.7 ± 7.8	88.5 ± 3.4	77.8 ± 2.7	55.8 ± 4.6	75.2 ± 11.3
BKM	05.0 ± 0.0	20.0 ± 0.1	10.8 ± 0.0	68.5 ± 0.0	07.7 ± 0.1	53.1 ± 2.6
SOBG	30.4 ± 0.0	16.8 ± 0.0	47.4 ± 0.0	84.2 ± 0.0	-	50.1 ± 0.0
NMTFCoS	63.2 ± 1.1	63.3 ± 1.8	86.7 ± 6.2	74.1 ± 9.4	59.2 ± 3.7	63.9 ± 3.2
MultiCC	65.0 ± 0.0	64.7 ± 0.0	84.0 ± 0.0	50.7 ± 0.0	60.6 ± 0.0	55.5 ± 0.0

Table 3 shows that RWGLCC has significantly higher clustering accuracy than others, such as USPS, LUNG and BASEHOCK. Since ISOLET is too sparse, it is not suitable for optimal bipartite graph methods such as SOBG. As shown in Table 4, taking NMI as the measurement standard, RWGLCC is slightly inferior to GNMF on ISOLET, but it still performs well. It can be seen from Table 5 that RWGLCC still achieves considerable performance with ARI as an indicator. It is worth mentioning that the performance of RWGLCC on BASEHOCK is far better than other algorithms. Besides, RWGLCC runs slightly longer than SNMF, roughly the same as GNMF and DNMTF, and much lower than others, reported in Fig. 4. In general, RWGLCC algorithm has certain competitiveness.

4.4. Ablation study

In this section, we consider several variants of our method. Variant 1 (V1) is the original non-negative matrix tri-factorization. Variant 2 (V2) replaces the F-norm of V1 with the 0-norm. Variant 3 (V3) introduces feature weights for V2. Variant 4 (V4) adds global discrimination to V3. Variant 5 (V5) attaches local discrimination to V4. It is evident from Table 6 that V5 with full components basically achieves the best performance on every dataset and metric, while V1 without any components achieves the lowest performance. Since USPS, JAFFE and COIL20 are relatively clean, the improvement of V2 is not obvious. Since ISOLET, LUNG, and BASE-

Table 4
Effectiveness test on multiple real datasets by normalized mutual information (NMI% ± std%).

Dataset	COIL20	USPS	JAFFE	LUNG	ISOLET	BASEHOCK
RWGLCC	**83.3±1.4**	**82.9±2.4**	**96.4±1.2**	**74.9±1.8**	74.6 ± 1.1	**58.9±1.6**
SNMF	74.7 ± 1.4	60.2 ± 1.4	85.0 ± 6.4	54.9 ± 4.8	73.6 ± 1.4	06.5 ± 0.2
GNMF	78.5 ± 1.8	76.2 ± 5.2	88.6 ± 3.1	57.7 ± 6.4	**75.7±1.5**	04.2 ± 8.8
FNMTF	74.9 ± 1.2	50.2 ± 1.5	80.1 ± 3.9	06.3 ± 4.9	68.7 ± 1.5	00.4 ± 0.0
DNMTF	78.1 ± 2.0	67.4 ± 9.6	90.0 ± 1.5	47.9 ± 7.1	72.5 ± 2.7	51.8 ± 16.2
BKM	01.3 ± 0.0	07.2 ± 0.1	04.1 ± 0.0	04.8 ± 0.0	11.4 ± 1.4	00.5 ± 0.5
SOBG	36.6 ± 0.0	00.3 ± 0.0	44.0 ± 0.0	58.8 ± 0.0	-	00.3 ± 0.0
NMTFCoS	74.0 ± 0.6	58.0 ± 0.7	90.1 ± 3.0	54.2 ± 7.4	73.9 ± 2.2	06.1 ± 3.0
MultiCC	75.2 ± 0.0	55.0 ± 0.0	84.8 ± 0.0	29.1 ± 0.0	68.6 ± 0.0	00.9 ± 0.0

Table 5
Effectiveness test on multiple real datasets by adjusted rand index (ARI% ± std%).

Dataset	COIL20	USPS	JAFFE	LUNG	ISOLET	BASEHOCK
RWGLCC	**68.4 ± 4.3**	**77.2 ± 7.2**	**94.8 ± 1.6**	**83.0 ± 1.4**	**55.3 ± 1.1**	**69.0 ± 1.4**
SNMF	57.1 ± 2.9	52.6 ± 1.4	74.4 ± 10.9	50.1 ± 10.7	51.8 ± 2.4	08.9 ± 0.2
GNMF	53.8 ± 4.9	62.4 ± 8.7	81.2 ± 5.8	62.2 ± 8.0	50.5 ± 4.6	02.8 ± 8.3
FNMTF	57.1 ± 2.3	41.4 ± 2.7	66.2 ± 6.4	00.6 ± 1.9	44.6 ± 1.9	00.5 ± 0.0
DNMTF	62.0 ± 4.2	57.9 ± 11.5	83.1 ± 2.4	44.8 ± 6.7	50.8 ± 4.4	47.5 ± 5.6
BKM	00.0 ± 0.0	00.2 ± 0.0	00.0 ± 0.0	00.0 ± 0.0	03.7 ± 0.8	00.6 ± 0.7
SOBG	07.4 ± 0.0	00.0 ± 0.0	13.7 ± 0.0	58.0 ± 0.0	-	00.0 ± 0.0
NMTFCoS	56.9 ± 0.9	50.2 ± 0.7	82.6 ± 6.1	47.3 ± 14.2	51.9 ± 3.1	08.0 ± 4.0
MultiCC	59.1 ± 0.0	47.3 ± 0.0	75.6 ± 0.0	18.2 ± 0.0	48.1 ± 0.0	01.2 ± 0.0

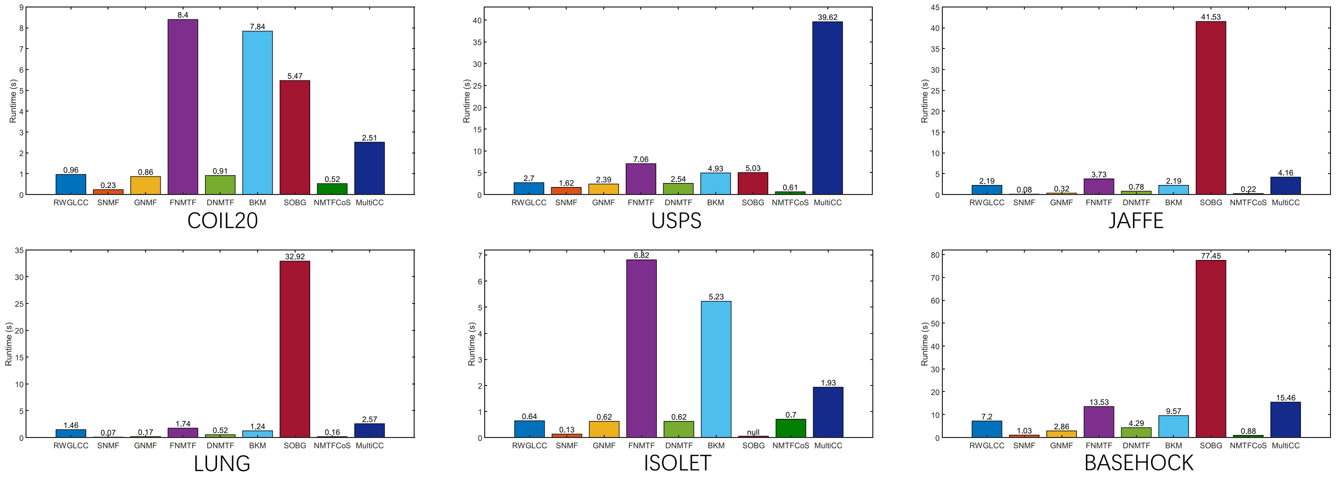


Fig. 4. Runtime for algorithm comparison on multiple real datasets.

HOCK are highly sparse, the performance of V2 even drops slightly. But a previous robustness test has shown it works. Compared with V1 and V2, V3 has a slight improvement. Compared with V3, V4 has obvious improvement, especially on COIL20, LUNG and BASEHOCK. Furthermore, the V5 has the most significant improvement over the other variants. It is clear that our proposed method is effective and its core components lie in global and local discrimination.

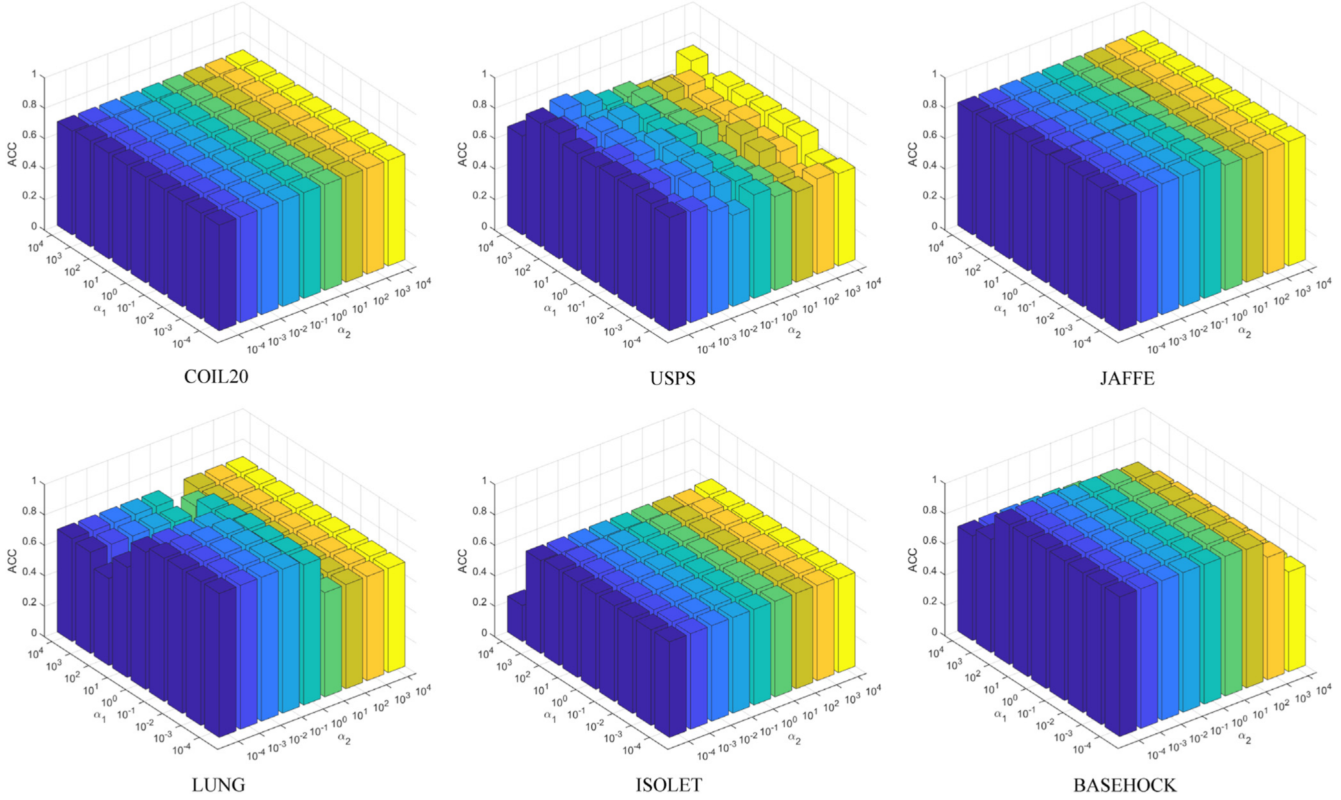
Local discrimination pays more attention to the nearest neighbor relationship of samples. Obviously, the closer the samples are, the more likely they are to belong to the same category. Global discrimination starts from the whole, requiring compact clusters and separation between clusters. This is also in line with the purpose of clustering. Therefore, using only local or global discrimination can improve performance. Since local discrimination essentially keeps the low-dimensional manifold in high-dimensional data, and better reveals the internal structure, it is generally better than global discrimination. However, if the boundary samples of two clusters are very close, the local relationship is harmful. At this time, the global discrimination plays a certain role in correction. Therefore, it is necessary to use both.

4.5. Parameter study

In this section, we analyze the performance of our algorithm as the parameters change. The influence of global and local discriminative regularizers is shown in Figs. 5 and 6, measured by clustering accuracy. From the results, local discrimination greatly impacts performance, and a relatively large value should be set, but too large a value will also cause performance degradation. The global discrimination usually plays a corrective role when the algorithm relies too much on the local structure. The performance is stable or even slightly improved at a smaller value, while a larger value will hinder the local structure.

4.6. Hypergraph discussion

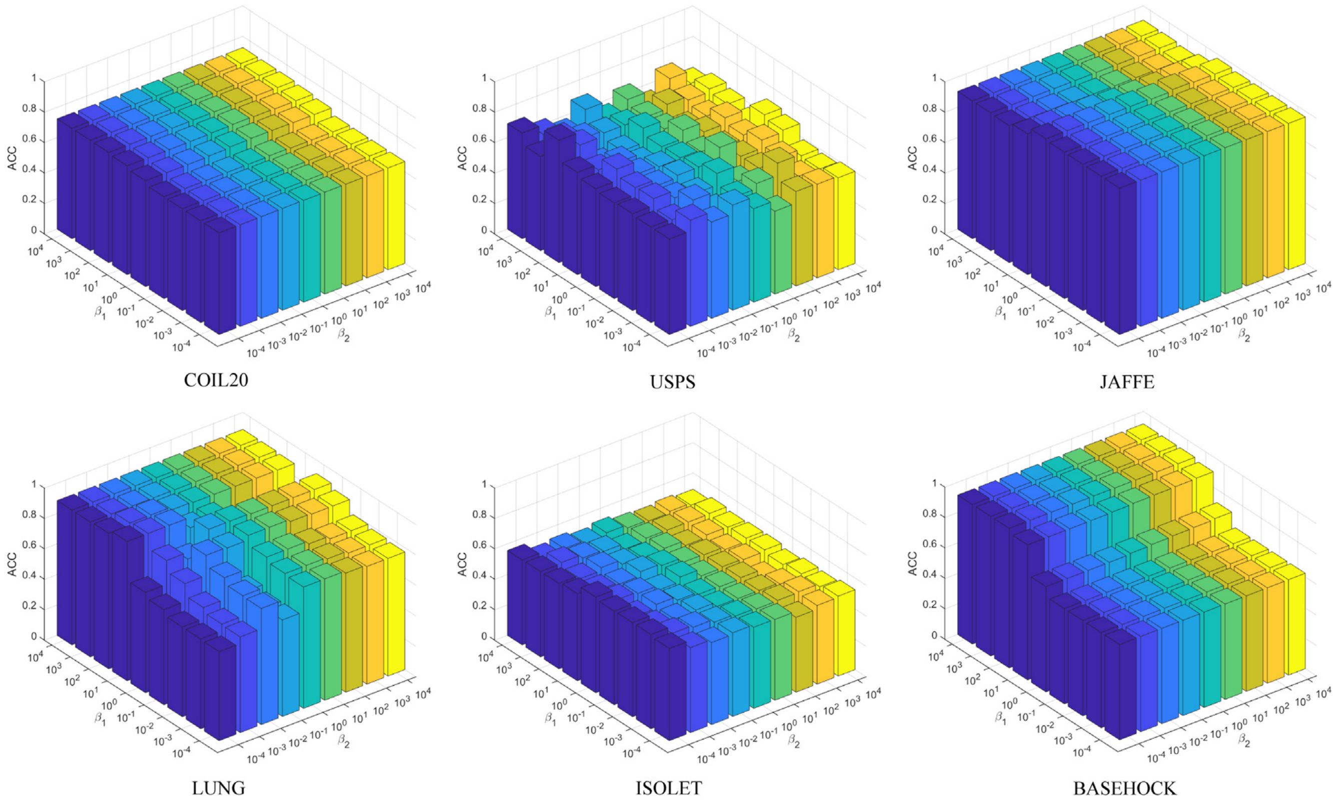
In many machine learning tasks, maintaining a local data structure is a common and effective way, usually in the form of graphs. In this paper, the n-ary relation among data is expressed as hypergraphs, and good results have been achieved. As shown in Table 7, the effect of hypergraph is usually better than that of traditional graph structure. Especially for BASEHOCK dataset, the effect of hy-



COIL20 USPS JAFFE

LUNG ISOLET BASEHOCK

Fig. 5. Accuracy variations with parameters α_1 and α_2.



COIL20 USPS JAFFE

LUNG ISOLET BASEHOCK

Fig. 6. Accuracy variations with parameters β_1 and β_2.

Table 6
Performance of several variants on different datasets and metrics.

Dataset	Metric	V1	V2	V3	V4	V5
	ACC	66.98	67.23	67.54	68.53	**80.92**
USPS	NMI	60.84	61.46	61.40	62.23	**82.88**
	ARI	52.87	53.49	53.53	55.05	**77.19**
	ACC	85.45	86.85	86.38	87.32	**97.56**
JAFFE	NMI	88.58	85.90	87.22	89.40	**96.39**
	ARI	80.17	77.39	80.48	81.97	**94.79**
	ACC	62.50	63.68	64.79	70.42	**74.42**
COIL20	NMI	74.34	74.41	74.58	77.12	**83.29**
	ARI	56.09	55.78	57.58	62.47	**68.38**
	ACC	58.91	57.50	60.83	62.56	**65.21**
ISOLET	NMI	73.42	**74.64**	74.63	73.08	74.55
	ARI	52.74	52.60	55.28	53.75	**55.34**
	ACC	61.58	60.10	61.58	70.44	**93.79**
LUNG	NMI	46.96	46.50	46.96	52.78	**74.86**
	ARI	32.48	32.36	32.48	42.28	**82.95**
	ACC	62.72	62.67	65.73	70.05	**91.53**
BASEHOCK	NMI	4.93	4.88	7.75	12.03	**58.87**
	ARI	6.43	6.38	9.85	16.03	**68.98**

pergraph is far better than that of graph. When constructing a hypergraph, a similarity between data still needs to be established, similar to a graph. Generally speaking, binary weights are easier to construct and calculate, and can often achieve good performance. Gaussian and cosine similarity more accurately preserve the differences between data, so they have better performance. In addition, Gaussian similarity is more suitable for image data, and cosine similarity is more suitable for text, but sometimes it is not.

5. Conclusion and future work

This paper mainly proposes a robust weighted co-clustering method with global discrimination and local discrimination. Specifically, in order to avoid the undesirable effects of data noise and consider the difference in feature importance, this method simultaneously models and learns data noise and feature weights during the clustering process. In addition, it considers the scatter from a global perspective and also considers the high-dimensional adjacency relationship from a local perspective. Finally, this paper discusses the algorithm's convergence and parameter sensitivity, and clarifies its excellent performance through a complete experiment. Currently, the emergence of multi-modal and multi-view data has promoted progress in various fields [43–45]. In future work, co-clustering should also explore more effective ways of using multi-view data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Proof of Theorem 1

Proof. First, rewrite $J(F)$ into the following form:

$$\begin{aligned} J(F) = &-\operatorname{Tr}\left(F^T M^+\right) + \operatorname{Tr}\left(F^T M^-\right) + \tfrac{1}{2}\operatorname{Tr}\left(F^T DFN^+\right) \\ &-\tfrac{1}{2}\operatorname{Tr}\left(F^T DFN^-\right) + \tfrac{1}{2}\operatorname{Tr}\left(F^T Q_F^+ F\right) - \tfrac{1}{2}\operatorname{Tr}\left(F^T Q_F^- F\right) \end{aligned} \tag{A.1}$$

Subject to Lemma 2, we obtain

$$\operatorname{Tr}(F^T DFN^+) \le \sum_{i=1}^{d}\sum_{j=1}^{c_1} \frac{\left(DF'N^+\right)_{ij} F_{ij}^2}{F'} \tag{A.2}$$

$$\operatorname{Tr}(F^T Q_F^+ F) \le \sum_{i=1}^{d}\sum_{j=1}^{c_1} \frac{(Q_F^+ F')_{ij} F_{ij}^2}{F'_{ij}} \tag{A.3}$$

Abiding by the inequality: $\forall a, b > 0, a \le \frac{a^2+b^2}{2b}$, we observe

$$\operatorname{Tr}(F^T M^-) = \sum_{i=1}^{d}\sum_{j=1}^{c_1} M_{ij}^- F_{ij} \le \sum_{i=1}^{d}\sum_{j=1}^{c_1} M_{ij}^- \frac{F_{ij}^2 + F_{ij}'^2}{2F'_{ij}} \tag{A.4}$$

Complied with the inequality: $\forall z > 0, z \ge 1 + \log z$, we notice

$$\begin{aligned} \operatorname{Tr}(F^T DFN^-) &= \textstyle\sum_{i=1}^{d}\sum_{j=1}^{c_1}\sum_{k=1}^{c_1} N_{jk}^- D_{ii} F'_{ij} F'_{ik} \frac{F_{ij}F_{ik}}{F'_{ij}F'_{ik}} \\ &\ge \textstyle\sum_{i=1}^{d}\sum_{j=1}^{c_1}\sum_{k=1}^{c_1} N_{jk}^- D_{ii} F'_{ij} F'_{ik} (1 + \log \frac{F_{ij}F_{ik}}{F'_{ij}F'_{ik}}) \end{aligned} \tag{A.5}$$

Table 7
Performance under different graph and hypergraph construction.

Dataset	Metric	Graph			Hypergraph		
		Binary	Cosine	Gaussian	Binary	Cosine	Gaussian
	ACC	75.53	75.17	75.22	75.25	**79.41**	79.21
USPS	NMI	76.90	75.51	75.55	80.67	**83.40**	82.24
	ARI	67.62	66.31	66.40	71.62	**75.61**	74.96
	ACC	85.45	86.85	87.79	96.24	95.77	**98.12**
JAFFE	NMI	87.51	89.04	90.65	94.21	93.39	**97.31**
	ARI	79.50	82.84	84.71	91.85	90.96	**95.92**
	ACC	67.01	65.49	65.69	73.40	73.47	**75.83**
COIL20	NMI	75.97	75.65	74.90	82.80	**83.44**	82.37
	ARI	61.59	59.40	59.39	68.98	70.21	**70.58**
	ACC	**65.38**	63.46	62.18	63.59	65.26	**65.38**
ISOLET	NMI	74.89	74.53	73.72	76.62	77.10	**77.29**
	ARI	56.53	54.93	54.91	**58.13**	57.93	57.45
	ACC	83.25	84.73	85.22	91.63	91.63	**93.60**
LUNG	NMI	64.73	55.35	56.28	69.26	69.26	**74.07**
	ARI	68.78	58.69	59.85	76.92	76.92	**82.32**
	ACC	74.06	73.31	73.96	89.96	88.51	**91.72**
BASEHOCK	NMI	17.72	16.50	17.51	53.57	46.77	**59.59**
	ARI	23.12	21.69	22.92	63.87	57.25	**69.61**

$$\begin{aligned}\mathrm{Tr}(F^T M^+) &= \sum_{i=1}^{d}\sum_{j=1}^{c_1} M^+_{ij}F_{ij} = \sum_{i=1}^{d}\sum_{j=1}^{c_1} M^+_{ij}F'_{ij}\frac{F_{ij}}{F'_{ij}} \\ &\geq \sum_{i=1}^{d}\sum_{j=1}^{c_1} M^+_{ij}F'_{ij}(1+\log\frac{F_{ij}}{F'_{ij}})\end{aligned} \tag{A.6}$$

$$\begin{aligned}\mathrm{Tr}(F^T Q^-_F F) &= \sum_{i=1}^{d}\sum_{j=1}^{c_1}\sum_{k=1}^{c_1} (Q^-_F)_{jk}F'_{ij}F'_{ik}\frac{F_{ij}F_{ik}}{F'_{ij}F'_{ik}} \\ &\geq \sum_{i=1}^{d}\sum_{j=1}^{c_1}\sum_{k=1}^{c_1} (Q^-_F)_{jk}F'_{ij}F'_{ik}(1+\log\frac{F_{ij}F_{ik}}{F'_{ij}F'_{ik}})\end{aligned} \tag{A.7}$$

Evidently, $Z(F,F')$ is an auxiliary function for $J(F)$, because $Z(F,F')$ is composed of all bounds to satisfy Definition 1.

To minimize $Z(F,F')$, we take

$$\begin{aligned}\frac{\partial Z(F,F')}{\partial F_{ij}} &= -\frac{M^+_{ij}F'ij}{F_{ij}} + \frac{M^-_{ij}F_{ij}}{F'ij} + \frac{(DF'\mathrm{N}^+)_{ij}F_{ij}}{F'_{ij}} \\ &-\frac{(DF'\mathrm{N}^-)_{ij}F'_{ij}}{F_{ij}} + \frac{(Q^+_F F')_{ij}F_{ij}}{F'_{ij}} - \frac{(Q^-_F F')_{ij}F'_{ij}}{F_{ij}}\end{aligned} \tag{A.8}$$

and obtain its Hessian matrix

$$\begin{aligned}\frac{\partial^2 Z(F,F')}{\partial F_{ij}\partial F_{kl}} &= \delta_{ik}\delta_{jl}\Bigg(\frac{M^+_{ij}F'i_{ij}}{F^2_{ij}} + \frac{M^-_{ij}}{F'_{ij}} + \frac{(DF'\mathrm{N}^+)_{ij}}{F'_{ij}} \\ &+\frac{(DF'\mathrm{N}^-)_{ij}F'_{ij}}{F^2_{ij}} + \frac{(Q^+_F F')_{ij}}{F'_{ij}} + \frac{(Q^-_F F')_{ij}F'_{ij}}{F^2_{ij}}\Bigg)\end{aligned} \tag{A.9}$$

where $\delta_{ik}=1$ if and only if $i=k$, otherwise, $\delta_{ik}=0$. The Hessian matrix is a non-negative diagonal matrix, hence $Z(F,F')$ is convex and $\arg\min_{F_{ij}} J(F) = \arg\min_{F_{ij}} Z(F,F')$ by setting $\frac{\partial Z(F,F')}{\partial F_{ij}} = 0$. □

References

[1] M. Rege, M. Dong, F. Fotouhi, Co-clustering documents and words using bipartite isoperimetric graph partitioning, in: Proceedings of the IEEE International Conference on Data Mining, 2006, pp. 532–541.

[2] R. Boutalbi, L. Labiod, M. Nadif, Sparse tensor co-clustering as a tool for document categorization, in: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 1157–1160.

[3] M. Selosse, J. Jacques, C. Biernacki, Textual data summarization using the self--organized co-clustering model, Pattern Recognit. 103 (2020) 107315.

[4] I. Konstas, V. Stathopoulos, J.M. Jose, On social networks and collaborative recommendation, in: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2009, pp. 195–202.

[5] E. Lima, W. Shi, X. Liu, Q. Yu, Integrating multi-level tag recommendation with external knowledge bases for automatic question answering, ACM Trans. Internet Technol. 19 (3) (2019) 1–22.

[6] L. Feng, Q. Zhao, C. Zhou, Improving performances of top-n recommendations with co-clustering method, Expert Syst. Appl. 143 (2020) 113078.

[7] M. Brameier, C. Wiuf, Co-clustering and visualization of gene expression data and gene ontology terms for saccharomyces cerevisiae using self-organizing maps, J. Biomed. Inform. 40 (2) (2007) 160–173.

[8] Y. Liu, Q. Gu, J.P. Hou, J. Han, J. Ma, A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression, BMC Bioinformatics 15 (37) (2014) 1–11.

[9] T. George, S. Merugu, A scalable collaborative filtering framework based on co-clustering, in: Proceedings of the IEEE International Conference on Data Mining, 2005, pp. 625–628.

[10] M. Khoshneshin, W.N. Street, Incremental collaborative filtering via evolutionary co-clustering, in: Proceedings of the ACM Conference on Recommender Systems, 2010, pp. 325–328.

[11] Y. Chen, M. Dong, W. Wan, Image co-clustering with multi-modality features and user feedbacks, in: Proceedings of the ACM International Conference on Multimedia, 2009, pp. 689–692.

[12] S.N. Vitaladevuni, R. Basri, Co-clustering of image segments using convex optimization applied to em neuronal reconstruction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2203–2210.

[13] S. Khan, L. Chen, H. Yan, Co-clustering to reveal salient facial features for expression recognition, IEEE Trans. Affect. Comput. 11 (2) (2020) 348–360.

[14] E. Giannakidou, V. Koutsonikola, A. Vakali, Y. Kompatsiaris, Co-clustering tags and social data sources, in: Proceedings of the International Conference on Web-Age Information Management, 2008, pp. 317–324.

[15] B.-K. Bao, W. Min, K. Lu, C. Xu, Social event detection with robust high-order co-clustering, in: Proceedings of the ACM International Conference on Multimedia Retrieval, 2013, pp. 135–142.

[16] G. Pio, M. Ceci, C. Loglisci, D. D'Elia, D. Malerba, Hierarchical and overlapping co-clustering of mRNA: miRNA interactions, in: Proceedings of the European Conference Artificial Intelligence, 2012, pp. 654–659.

[17] J. Luo, B. Liu, B. Cao, S. Wang, Identifying miRNA-mRNA regulatory modules based on overlapping neighborhood expansion from multiple types of genomic data, in: Proceedings of the International Conference on Intelligent Computing Theories and Application,, 2016, pp. 234–246.

[18] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (6755) (1999) 788–791.

[19] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, in: Proceedings of the Annual Conference on Neural Information Processing Systems, 2000, pp. 556–562.

[20] C. Ding, X. He, H.D. Simon, On the equivalence of nonnegative matrix factorization and spectral clustering, in: Proceedings of the SIAM International Conference on Data Mining, 2005, pp. 606–610.

[21] T. Li, C. Ding, The relationships among various nonnegative matrix factorization methods for clustering, in: Proceedings of the IEEE International Conference on Data Mining, 2006, pp. 362–371.

[22] C. Ding, T. Li, M. Jordan, Convex and semi-nonnegative matrix factorizations, IEEE Trans. Pattern Anal. Mach. Intell. 32 (1) (2010) 45–55.

[23] D. Cai, X. He, J. Han, T.S. Huang, Graph regularized nonnegative matrix factorization for data representation, IEEE Trans. Pattern Anal. Mach. Intell. 33 (8) (2011) 1548–1560.

[24] Q. Gu, J. Zhou, Co-clustering on manifolds, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009, pp. 359–368.

[25] F. Shang, L. Jiao, F. Wang, Graph dual regularization non-negative matrix factorization for co-clustering, Pattern Recognit. 45 (6) (2012) 2237–2250.

[26] C. Ding, T. Li, W. Peng, H. Park, Orthogonal nonnegative matrix t-factorizations for clustering, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006, pp. 126–135.

[27] S. Wang, W. Guo, Robust co-clustering via dual local learning and high-order matrix factorization, Knowl. Based Syst. 138 (2017) 176–187.

[28] S. Wang, A. Huang, Penalized nonnegative matrix tri-factorization for co-clustering, Expert Syst. Appl. 78 (2017) 64–73.

[29] H. Wang, F. Nie, H. Huang, F. Makedon, Fast nonnegative matrix tri-factorization for large-scale data co-clustering, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2011, pp. 1553–1558.

[30] Q. Tan, P. Yang, J. He, Feature co-shrinking for co-clustering, Pattern Recognit. 77 (2018) 12–19.

[31] J. Wang, X. Wang, G. Yu, C. Domeniconi, Z. Yu, Z. Zhang, Discovering multiple co-clusterings with matrix factorization, IEEE Trans. Cybern. 51 (7) (2021) 3576–3587.

[32] L. Zhang, C. Chen, J. Bu, Z. Chen, D. Cai, J. Han, Locally discriminative coclustering, IEEE Trans. Knowl. Data Eng. 24 (6) (2011) 1025–1035.

[33] J. Han, K. Song, F. Nie, X. Li, Bilateral k-means algorithm for fast co-clustering, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2017, pp. 1969–1975.

[34] F. Nie, X. Wang, C. Deng, H. Huang, Learning a structured optimal bipartite graph for co-clustering, in: Proceedings of the Annual Conference on Neural Information Processing Systems, 2017, pp. 4129–4138.

[35] I.S. Dhillon, S. Mallela, D.S. Modha, Information-theoretic co-clustering, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003, pp. 89–98.

[36] X. Yu, G. Yu, J. Wang, C. Domeniconi, Co-clustering ensembles based on multiple relevance measures, IEEE Trans. Knowl. Data Eng. 33 (4) (2021) 1389–1400.

[37] S.F. Hussain, A. Pervez, M. Hussain, Co-clustering optimization using Artificial Bee Colony (ABC) algorithm, Appl. Soft Comput. 97 (2020) 106725.

[38] J. Wright, A. Ganesh, S.R. Rao, Y. Peng, Y. Ma, Robust principal component analysis: exact recovery of corrupted low-rank matrices via convex optimization, in: Proceedings of the Annual Conference on Neural Information Processing Systems, Vol. 58, 2009, pp. 289–298.

[39] J. Ye, Z. Zhao, M. Wu, Discriminative k-means for clustering, in: Proceedings of the Annual Conference on Neural Information Processing Systems, Vol. 20, 2007, pp. 1649–1656.

[40] Y. Yang, H.T. Shen, F. Nie, R. Ji, X. Zhou, Nonnegative spectral clustering with discriminative regularization, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2011, pp. 555–560.

[41] J.A. Hartigan, M.A. Wong, A k-means clustering algorithm, J. R. Stat. Soc. 28 (1) (1979) 100–108.

[42] L. van der Maaten, G. Hintton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (1) (2008) 2579–2625.

[43] S. Huang, Z. Xu, I.W. Tsang, Z. Kang, Auto-weighted multi-view co-clustering with bipartite graphs, Inf. Sci. 512 (2020) 18–30.

[44] F. Nie, S. Shi, X. Li, Auto-weighted multi-view co-clustering via fast matrix factorization, Pattern Recognit. 102 (2020) 107207.

[45] X. Yan, Z. Lou, S. Hu, Y. Ye, Multi-task information bottleneck co-clustering for unsupervised cross-view human action categorization, ACM Trans. Knowl. Discov. Data 14 (2) (2020) 1–23.

Zhoumin Lu received the MS degree in computer technology from Fuzhou University, China, in 2021. He is currently pursuing the PhD degree with the School of Computer Science, Northwestern Polytechnical University. His research interests include machine learning and its applications, such as pattern recognition and data mining.

Shiping Wang received the PhD degree from the School of Computer Science and Engineering, University of Electronic Science and Technology of China, China, in 2014. He is currently a Full Professor with the College of Mathematics and Computer Science at Fuzhou University. His research interests include machine learning and computer vision.

Genggeng Liu received the PhD degree in Applied Mathematics from Fuzhou University, China, in 2015. He is currently an Associate Professor with the College of

Computer and Data Science at Fuzhou University. His research interests include computational intelligence and very large scale integration physical design.

Feiping Nie received the PhD degree in computer science from Tsinghua University, China, in 2009. He is currently a Full Professor with Northwestern Polytechnical University. He has published more than 100 articles in the following journals and conferences TPAMI, IJCV, TIP, TNNLS, TKDE, ICML, NIPS, KDD, IJCAI, AAAI, ICCV, CVPR, and ACM MM. His articles have been cited more than 10000 times and the H-index is 57. His research interests include machine learning and its applications, such as pattern recognition, data mining, computer vision, image processing, and information retrieval. He is currently serving as a PC member or an associate editor for several prestigious journals and conferences in the related fields.