Tsinghua Science and Technology

Volume 26 | Issue 2 Article 6

2021

TW-Co-MFC: Two-Level Weighted Collaborative Fuzzy Clustering Based on Maximum Entropy for Multi-View Data

Jie Hu
School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China

Yi Pan
Department of Computer Science, Georgia State University, Atlanta, GA 30302-3994, USA.

Tianrui Li
School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China

Yan Yang
School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China

Recommended Citation

Jie Hu, Yi Pan, Tianrui Li et al. TW-Co-MFC: Two-Level Weighted Collaborative Fuzzy Clustering Based on Maximum Entropy for Multi-View Data. Tsinghua Science and Technology 2021, 26(2): 185-198.

TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 06/11 pp185–198
DOI: 10.26599/TST.2019.9010078
Volume 26, Number 2, April 2021

TW-Co-MFC: Two-Level Weighted Collaborative Fuzzy Clustering Based on Maximum Entropy for Multi-View Data

Jie Hu*, Yi Pan, Tianrui Li, and Yan Yang

Abstract: In recent years, multi-view clustering research has attracted considerable attention because of the rapidly growing demand for unsupervised analysis of multi-view data in practical applications. Despite the significant advances in multi-view clustering, two challenges still need to be addressed, i.e., how to make full use of the consistent and complementary information in multiple views and how to discriminate the contributions of different views and features in the same view to efficiently reveal the latent cluster structure of multi-view data for clustering. In this study, we propose a novel Two-level Weighted Collaborative Multi-view Fuzzy Clustering (TW-Co-MFC) approach to address the aforementioned issues. In TW-Co-MFC, a two-level weighting strategy is devised to measure the importance of views and features, and a collaborative working mechanism is introduced to balance the within-view clustering quality and the cross-view clustering consistency. Then an iterative optimization objective function based on the maximum entropy principle is designed for multi-view clustering. Experiments on real-world datasets show the effectiveness of the proposed approach.

Key words: multi-view clustering; fuzzy clustering; collaborative; weighting; maximum entropy

1 Introduction

Multi-view data are nearly omnipresent in real-world applications, because large amounts of data are collected from diverse domains using different measurement methods and represented by different feature groups that contain specific information about these data[1,2]. For instance, in image analysis, images can be represented by local shape, color descriptors, and local binary patterns. In web page classification, web pages consist of text content, embedded pictures, and hyperlink connecting to other pages. With the goal of revealing the latent group structures in multi-view datasets, multi-view clustering has become a promising direction in machine learning and numerous multi-view clustering methods have been proposed.

Generally, on the basis of the view fusion position during the clustering process, view fusion strategies can be classified into three types, namely, prior fusion, posterior fusion, and middle fusion methods[2,3]. The prior fusion method usually refers to the direct concatenation of each feature of all views into one unique representation or the indirect addition of the similarity matrix derived from each view and the use of single-view clustering algorithms to obtain the final result[4,5]. However, this kind of view fusion method does not improve the clustering performance, because the concatenation operation leads to not only the overfitting phenomenon on a small training sample but also the loss of specific statistical properties and complementary nature of different views. In addition, the concatenation operation may also lead to the dimensionality problem. The posterior fusion method is also known as the clustering ensemble method, which

• Jie Hu, Tianrui Li, and Yan Yang are with the School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China. E-mail: jiehu@swjtu.edu.cn; trli@swjtu.edu.cn; yyang@swjtu.edu.cn.
• Yi Pan is with Department of Computer Science, Georgia State University, Atlanta, GA 30302-3994, USA. E-mail: yipan@gsu.edu.
∗ To whom correspondence should be addressed.
Manuscript received: 2019-12-08; accepted: 2019-12-15

is applied to achieve consensus clustering assignment among the set of clustering results obtained from the individual clustering process of each view[6]. The main shortcoming of this approach is the lack of consideration of the complementary and consistent information between views in the previous base clustering phase, which may result in an unstable performance.

By contrast, the middle fusion method uses information from various views simultaneously in the clustering process to obtain the clustering results. On the basis of the specific fusion mechanism of multiple views during the clustering process, middle fusion methods can be divided into two categories. The first category considers the hypothesis that the clustering results in different views should be unique[7–12]. Therefore, in the execution process of a multi-view cluster, the multi-view information acts on common and consistent results sequentially. During this process, nearly no interaction occurs between views. The second category involves clustering the information in each view while collaboratively integrating the information from other views and adopting the clustering ensemble method to combine the previous single-view clustering results to obtain a final consensus result[3,5,13–17]. In these approaches, the consistency between views and the uniqueness of each view are both considered. For example, Jiang et al.[5] proposed a weighted view collaborative Fuzzy c-Means (FCM), which incorporated two penalty terms to improve the result, where the first term was designed to automatically reward or suppress the corresponding fuzzy membership degrees of a fixed sample point for the current view and the other views and the second term was designed to reduce the differences of different view partitioning results in a collaborative manner. On the basis of the max-product belief propagation, Wang et al.[14] established a collaborative multi-view clustering objective function consisting of two components, in which one component was used to measure the within-view clustering quality and the other component was introduced to assess the explicit clustering consistency across different views. Zeng et al.[17] proposed a unified collaborative multi-kernel fuzzy clustering for multi-view data by unifying the local partitions and global clustering result in a collaborative learning framework. Firstly, all of the features were projected on the common multi-kernel space. Then a single fuzzy clustering objective function with two parts, i.e., one part was the joint local partition clustering method and the other part was the classic weighted multi-kernel clustering method, was proposed. Although the proposed algorithm showed excellent performance, the computational cost of constructing the combined kernel expression for each view was relatively large and remarkable. Given that collaborative multi-view clustering not only fully considers the characteristics of target objects in multiple angles, but also takes full advantage of the consistency between all of the views involved, the final decision is usually more reliable than that of the other two strategies in terms of the principle of retaining differences while seeking common ground.

In addition, the importance of each view should also be assessed in real clustering applications, because the view of data has either strong or weak discrimination capability in essence, and some views acquired by unreliable physical measurements may be even damaged by noise. Several scholars addressed this issue and proposed the corresponding strategies[5,17–20]. For example, Xia et al.[18] presented a robust multi-view spectral clustering method that learned a shared low-rank transition probability matrix by separating noise from each graph and utilized the classic Markov chain method for clustering. Jiang et al.[5] developed a collaborative fuzzy clustering algorithm, which introduced a term to punish the single sample point that had an ambiguous membership degree upon clustering and added a weight parameter to measure the importance of various views. Xu et al.[20] proposed a re-weighted discriminative embedded k-means, which can effectively mitigate the influence of outliers and achieve dimension reduction while adaptively weighting diverse views using a multi-view least absolute residual model. However, their work only considered the importance of views, not the importance of different features in the same view, which may result in a significant decrease in clustering performance when the uncorrelated, interactive, redundant or noisy features were blended in the feature space. Therefore, several researchers proposed bi-level weighting multi-view clustering methods. For example, Chen et al.[8] presented a strategy of simultaneous weighting of views and features to accomplish a multi-view clustering task under the classical k-means framework. They used a fuzzy weighting strategy to represent the importance of views and features. A similar idea was adopted by Jiang et al.[9] Given that the high dimensionality of features may lead to a high-complexity and low-stability clustering performance, Xu et al.[10] proposed a multi-view clustering method with feature selection that can simultaneously provide the weights of views and features. Zhang et al.[15] presented a collaborative multi-view k-means clustering method, which considered the importance of views and features and worked in a collaborative manner to take advantage of the complementary and consistent information across different views.

Recently, fuzzy clustering analysis approaches, as important complex data analysis methods used to deal with uncertain, imprecise, and incomplete information, have triggered extensive research[21]. The representative works of fuzzy clustering analysis can be divided into four categories, namely, FCM and its derivatives, Maximum Entropy Clustering (MEC), hybrid rough-fuzzy clustering approaches, and other fuzzy clustering models as well as applications[22]. Although FCM has been widely considered as the most representative of fuzzy partition clustering methods, its weighted index that is used to generate the fuzzy membership function lacks physical meaning[23]. Therefore, the weighted index is neither mathematically natural nor necessary[23,24]. Different from FCM, MEC applies the maximum entropy principle to generate the fuzzy membership function. The parameter of MEC reflects a clear physical meaning by adding an entropy term to the objective function[23]. Many MEC-based fuzzy clustering methods have been proposed[22,23,25]. For instance, Qian et al.[26] proposed a multi-view MEC by jointly leveraging inter-view collaborations and intra-view- weighted attributes. However, their work has not taken into account the weighting of views.

The previously presented discussion indicates that the existing methods are far from being able to completely solve the problems encountered in multi-view cluster analysis. In this study, we propose a novel Two-level Weighted Collaborative Multi-view Fuzzy Clustering (TW-Co-MFC) approach based on maximum entropy to address the aforementioned issues. In TW-Co-MFC, a two-level weighting strategy, which is designed to simultaneously measure the importance of views and features in each view, is devised in an iterative re-weighted manner to examine the difference between views and the difference between features in the same view to improve the clustering performance. On the basis of the maximum entropy principle, an iterative optimization objective function is designed to balance the within-view clustering quality and the cross-view clustering consistency in a collaborative manner.

The main contributions of the proposed approach are summarized as follows:

- A new fuzzy multi-view clustering objective function is proposed under the MEC framework, which provides a good explanation for the fuzzy partition.
- Both the importance of views and features in each view are considered in the objective function.
- The consistent and complementary information between views are fully utilized in a collaborative manner.
- Extensive experiments on a variety of real-world datasets show the effectiveness of the proposed method.

The remainder of this paper is organized as follows: The related works on collaborative fuzzy clustering and MEC are outlined in Section 2. The basic idea, key steps, and proposed algorithm are described in Section 3. The experimental results and the corresponding parameter analysis are discussed in Section 4. The conclusions and future research topics are presented in Section 5.

2 Related Work

In this section, we briefly review collaborative fuzzy clustering and MEC.

2.1 Related notation

In this study, we consider a set of data $X = \{x_1, \ldots, x_i, \ldots, x_N\}$ with N samples, T views, and K clusters, where $1 \leqslant i \leqslant N$. For $1 \leqslant t \leqslant T$and $1 \leqslant k \leqslant K$, in the t-th view, G_t denotes the number of the features and $x_{ij,t}$ denotes the j-th feature value of the i-th sample in the t-th view with $j \in G_t$. In the fuzzy clustering framework, to cluster the sample set X into K classes, $u_{ik,t} \in [0, 1]$ denotes the fuzzy membership degree of x_i to cluster k in view t, $v_{k,t}$ denotes the center of cluster k in view t, $d_{ik,t}$ denotes the Euclidean distance between x_i and v_k in view t, $v_{kj,t}$ denotes the j-th feature value of k-th cluster center in t-th view, w_t represents the weight of t-th view, and $m_{j,t}$ denotes the j-th feature weight in the t-th view.

2.2 Collaborative fuzzy clustering: Co-FKM algorithm

FCM[27] is a well-known classical fuzzy clustering algorithm. Using the conventional FCM framework, Cleuziou et al.[28] developed a multi-view clustering approach, i.e., Co-FKM, which has become the basis of the collaborative multi-view fuzzy clustering method.

In Co-FKM, each view has a specific partition and a penalty term is introduced to reduce the inconsistency between partitions from different views. The objective function is defined as the minimization of the distance between samples and cluster centers in each view while penalizing the disagreement between any pairs of views.

$$J_{\text{Co-FKM}}(U,V)=\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{k=1}^{K}u_{ik,t}^{m}d_{ik,t}^{2}+\eta\frac{1}{T-1}\sum_{t'=1,t'\neq t}^{T}\sum_{i=1}^{N}\sum_{k=1}^{K}(u_{ik,t'}^{m}-u_{ik,t}^{m})d_{ik,t}^{2}=\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{k=1}^{\text{K}}\bar{u}_{ik,t,\eta}d_{ik,t}^{2},$$
$$\text{s.t.}\sum_{k=1}^{K}u_{ik,t}=1,\ 0\leqslant u_{ik,t}\leqslant 1, 1\leqslant i\leqslant N, 1\leqslant k\leqslant K, 1\leqslant t\leqslant T \tag{1}$$

where $\bar{u}_{ik,t,\eta}=(1-\eta)u_{ik,t}^{m}+\frac{\eta}{T-1}\sum_{t'=1,t'\neq t}^{T}u_{ik,t'}^{m}$ and η is a parameter used to control the penalty associated with the disagreement. The term $\frac{1}{T-1}\sum_{t'=1,t'\neq t}^{T}\sum_{i=1}^{N}\sum_{k=1}^{K}(u_{ik,t'}^{m}-u_{ik,t}^{m})d_{ik,t}^{2}$ is a disagreement term, which can be considered as the divergence between partitions from different views, i.e., the lower the value of $(u_{ik,t'}^{m}-u_{ik,t}^{m})$, the lower the disagreement.

In Co-FKM, the idea of clustering ensemble is adopted to combine individual view fuzzy partition $u_{ik,t}$ and obtain the global clustering result $\hat{u}_{ik}$. The consensus function is defined as the geometric mean of $u_{ik,t}$ for each view and expressed as follows:

$$\hat{u}_{ik}=\sqrt[T]{\prod_{t=1}^{T}u_{ik,t}} \tag{2}$$

Co-FKM improved the performance of multi-view clustering, and as indicated in Eq. (1), Co-FKM considered that each view and each feature contributed equally to clustering, which may decrease the clustering performance when the views and features had different importance.

2.3 MEC

Actually, MEC is a kind of fuzzy clustering method that includes some form of maximum entropy term in the objective function[22]. The most classic MEC model[23,29] can be expressed as follows:

$$\min_{V,U}\left(\sum_{k=1}^{K}\sum_{i=1}^{N}u_{ik}\|x_i-v_k\|^2+\beta\sum_{k=1}^{K}\sum_{i=1}^{N}u_{ik}\ln u_{ik}\right),$$
$$\text{s.t.}\sum_{k=1}^{K}u_{ik}=1\quad\text{and}\quad 0\leqslant u_{ik}\leqslant 1, 1\leqslant i\leqslant N, 1\leqslant k\leqslant K \tag{3}$$

where the fuzzy membership $\sum_{i,j}u_{ij}\ln u_{ij}$ is derived from Shannon entropy and β is the regularization parameter.

3 A Two-Level Weighted Collaborative Fuzzy Clustering Based on Maximum Entropy for Multi-View Data

In this section, firstly, the objective function of the proposed TW-Co-MFC is formulated. Then the updating rules are derived by applying the Lagrangian multiplier method to the iterative clustering process to minimize the objective function. Subsequently, the TW-Co-MFC algorithm and its steps are introduced in detail. Finally, the time complexity of the algorithm is discussed.

3.1 Objective function

The collaborative multi-view fuzzy clustering process to partition X into K clusters with weights for both views and individual features is modeled as the minimization of the following objective function:

$$J_{\text{TW-Co-MFC}}(U,V,W,M)=\sum_{t=1}^{T}w_t\left(\sum_{i=1}^{N}\sum_{k=1}^{K}\widetilde{u}_{ik,t,\eta}\sum_{j\in G_t}m_{j,t}\left(x_{ij,t}-v_{kj,t}\right)^2\right)+\lambda_1\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{k=1}^{K}u_{ik,t}\ln u_{ik,t}+\lambda_2\sum_{t=1}^{T}w_t\ln w_t+\lambda_3\sum_{t=1}^{T}\sum_{j\in G_t}m_{j,t}\ln m_{j,t} \tag{4}$$

s.t.

$$\begin{cases}\sum_{k=1}^{K}u_{ik,t}=1, u_{ik,t}\in[0,1],\ 1\leqslant i\leqslant N, 1\leqslant t\leqslant T;\\ \sum_{t=1}^{T}w_t=1, w_t\in[0,1];\\ \sum_{j\in G_t}m_{j,t}=1\ , m_{j,t}\in[0,1]\ , 1\leqslant t\leqslant T;\end{cases}$$

where $\widetilde{u}_{ik,t,\eta}$ is the weighted membership degree obtained from each view, which is defined as follows:

$$\widetilde{u}_{ik,t,\eta} = (1-\eta)u_{ik,t} + \frac{\eta}{T-1}\sum_{t'=1,t'\neq t}^{T} u_{ik,t'} \quad (5)$$

The parameter η is a trade-off factor used to control the penalty associated with the disagreement of cross-view clustering results. Three parameters, i.e., λ_1, λ_2, and λ_3, are set to control the distributions of fuzzy membership U, view weighting variable W, and feature weighting variable M, respectively.

The objective function expressed in Eq. (4) consists of four terms. The first term collaboratively measures the total deviation between all samples and all cluster centers in all views using the "with a two-level weighting strategy" (i.e., the first level is used to weight view importance and the second level is used to weight feature importance). The remaining terms, i.e., $\alpha\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{k=1}^{K} u_{ik,t}\ln u_{ik,t}$, $\beta\sum_{t=1}^{T} w_t \ln w_t$, and $\gamma\sum_{t=1}^{T}\sum_{j\in G_t} m_{j,t}\ln m_{j,t}$, are three maximum entropy-based terms adopted to search for the unbiased probability assignments for fuzzy membership, view weighting, and feature weighting throughout the clustering process, respectively.

To obtain the overall fuzzy partition matrix $\overline{U}$, the idea of clustering ensemble is adopted to calculate the summation of the weighted fuzzy partition matrix from each view.

$$\overline{U} = \sum_{t=1}^{T} w_t U_t \quad (6)$$

where U_t is the fuzzy membership matrix of the t-th view.

3.2 Clustering optimization

To minimize the objective function expressed in Eq. (4), we adopt an iterative optimization strategy that includes four steps. In each step, we determine the optimal variable by fixing the three other variables. Note that the variable optimization order does not affect the output. We decide to update cluster centroid V firstly just because we randomly select and initialize U, M, and W before the iteration. The optimizations of U, M, and W are independent. Under the convergence condition, the difference between the value of the objective function in current iteration and that of the last iteration is less than a minimal value like 10^{-6} or reaches the maximum iteration. The detailed description of the optimization process is provided in the following subsections.

3.2.1 Cluster centroid optimization

In this step, we update the cluster centroid of the j-th feature $v_{kj,t}$ by fixing variables U, W, and M in each view.

We seek to derive the optimal prototypes as follows. By setting $\partial J_{\text{TW-Co-MFC}}/\partial v_{kj,t} = 0$, we obtain the following expression:

$$\frac{\partial J_{\text{TW-Co-MFC}}}{\partial v_{kj,t}} = -2w_t\sum_{i=1}^{N}\widetilde{u}_{ik,t,\eta}m_{j,t}\left(x_{ij,t}-v_{kj,t}\right)=0 \quad (7)$$

$J_{\text{TW-Co-MFC}}$ reaches the local minimum if and only if $v_{kj,t}$ meets the following condition:

$$v_{kj,t} = \frac{\sum_{i=1}^{N}\widetilde{u}_{ik,t,\eta}m_{j,t}x_{ij,t}}{\sum_{i=1}^{N}\widetilde{u}_{ik,t,\eta}m_{j,t}} \quad (8)$$

3.2.2 Feature weight optimization

In this step, we update the feature weight of the j-th feature of the t-th view $v_{kj,t}$ by fixing variables U, W, and M in each view.

Through Lagrangian optimization, the minimum value of $J_{\text{TW-Co-MFC}}$ in Eq. (4) can be obtained by solving the following optimization problem about $m_{j,t}$:

$$\begin{aligned} &J_{\text{TW-Co-MFC}}\left(m_{j,t},\gamma_t\right) = \\ &\sum_{t=1}^{T} w_t\left(\sum_{i=1}^{N}\sum_{k=1}^{K}\widetilde{u}_{ik,t,\eta}\sum_{j\in G_t} m_{j,t}\left(x_{ij,t}-v_{kj,t}\right)^2\right)+ \\ &\lambda_1\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{k=1}^{K}u_{ik,t}\ln u_{ik,t} + \lambda_2\sum_{t=1}^{T}w_t\ln w_t + \\ &\lambda_3\sum_{t=1}^{T}\sum_{j\in G_t}m_{j,t}\ln m_{j,t} - \sum_{t=1}^{T}\gamma_t\left(\sum_{j\in G_t}m_{j,t}-1\right) \end{aligned} \quad (9)$$

where r_t is a Lagrange multiplier.

From Eq. (9), we obtain the optimal value of $m_{j,t}$ by setting $\partial J_{\text{TW-Co-MFC}}\left(m_{j,t},\gamma_t\right)/\partial m_{j,t} = 0$ and $\partial J_{\text{TW-Co-MFC}}\left(m_{j,t},\gamma_t\right)/\partial\gamma_t = 0$. Thus, we have

$$\begin{aligned} &\frac{\partial J_{\text{TW-Co-MFC}}\left(m_{j,t},\gamma_t\right)}{\partial m_{j,t}} = \\ &w_t\left(\sum_{i=1}^{N}\sum_{k=1}^{K}\widetilde{u}_{ik,t,\eta}\left(x_{ij,t}-v_{kj,t}\right)^2\right)+ \\ &\lambda_3\left(\ln m_{j,t}+1\right)-\gamma_t = 0 \end{aligned} \quad (10)$$

$$\frac{\partial J_{\text{TW-Co-MFC}}\left(m_{j,t},\gamma_t\right)}{\partial\gamma_t} = \sum_{j\in G_t} m_{j,t} - 1 = 0 \quad (11)$$

From Eq. (10), $m_{j,t}$ is acquired as follows:

$$m_{j,t} = \exp\left(\frac{\gamma_t + \lambda_3}{\lambda_3}\right) \times \exp\left(\frac{w_t \sum_{i=1}^{N}\sum_{k=1}^{K} \widetilde{u}_{ik,t,\eta}\left(x_{ij,t} - v_{kj,t}\right)^2}{-\lambda_3}\right) \quad (12)$$

With the constraint of Eq. (11), we have

$$\exp\left(\frac{\gamma_t + \lambda_3}{\lambda_3}\right) = \frac{1}{\sum_{s\in G_t} \exp\left(\frac{w_t \sum_{i=1}^{N}\sum_{k=1}^{K} \widetilde{u}_{ik,t,\eta}\left(x_{is,t} - v_{ks,t}\right)^2}{-\lambda_3}\right)} \quad (13)$$

Then by substituting Eq. (13) into Eq. (12), we acquire the solution of alternative optimal weight for the j-th feature in the t-th view as

$$m_{j,t} = \frac{\exp\left(\frac{\sum_{i=1}^{N}\sum_{k=1}^{K} \widetilde{u}_{ik,t,\eta}\left(x_{ij,t} - v_{kj,t}\right)^2}{-\lambda_3}\right)}{\sum_{j'\in G_t} \exp\left(\frac{\sum_{i=1}^{N}\sum_{k=1}^{K} \widetilde{u}_{ik,t,\eta}\left(x_{ij',t} - v_{kj',t}\right)^2}{-\lambda_3}\right)} \quad (14)$$

3.2.3 View weight optimization

In this step, we update the weight w_t of t-th view by fixing variables U, V, and M in each view.

By the Lagrangian optimization, the minimum of $J_{\text{TW-Co-MFC}}$ in Eq. (4) can be solved by finding the following:

$$J_{\text{TW-Co-MFC}}(w_t, \beta) = \sum_{t=1}^{T} w_t \left(\sum_{i=1}^{N}\sum_{k=1}^{K} \widetilde{u}_{ik,t,\eta} \sum_{j\in G_t} m_{j,t}\left(x_{ij,t} - v_{kj,t}\right)^2\right) + \lambda_1 \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{k=1}^{K} u_{ik,t} \ln u_{ik,t} + \lambda_2 \sum_{t=1}^{T} w_t \ln w_t + \lambda_3 \sum_{t=1}^{T}\sum_{j\in G_t} m_{j,t} \ln m_{j,t} - \beta\left(\sum_{t=1}^{T} w_t - 1\right) \quad (15)$$

From Eq. (15), we can obtain the optimal value of w_t by setting $\partial J_{\text{TW-Co-MFC}}(w_t, \beta)/\partial w_t = 0$ and $\partial J_{\text{TW-Co-MFC}}(w_t, \beta)/\partial \beta = 0$. Thus, we derive the following expressions:

$$\frac{\partial J_{\text{TW-Co-MFC}}(w_t, \beta)}{\partial w_t} = \sum_{i=1}^{N}\sum_{k=1}^{K} \widetilde{u}_{ik,t,\eta} \sum_{j\in G_t} m_{j,t}\left(x_{ij,t} - v_{kj,t}\right)^2 + \lambda_2 (\ln w_t + 1) - \beta = 0 \quad (16)$$

$$\frac{\partial J_{\text{TW-Co-MFC}}(w_t, \beta)}{\partial \beta} = \sum_{t=1}^{T} w_t - 1 = 0 \quad (17)$$

From Eq. (16), w_t is acquired as follows:

$$w_t = \exp\left(\frac{\beta + \lambda_2}{\lambda_2}\right) \times \exp\left(\frac{\sum_{i=1}^{N}\sum_{k=1}^{K} \widetilde{u}_{ik,t,\eta} \sum_{j\in G_t} m_{j,t}\left(x_{ij,t} - v_{kj,t}\right)^2}{-\lambda_2}\right) \quad (18)$$

Using the constraint of Eq. (17), we derive the following expression:

$$\exp\left(\frac{\beta + \lambda_2}{\lambda_2}\right) = \frac{1}{\sum_{r=1}^{T} \exp\left(\frac{\sum_{i=1}^{N}\sum_{k=1}^{K} \widetilde{u}_{ik,r,\eta} \sum_{j\in G_r} m_{j,r}\left(x_{ij,r} - v_{kj,r}\right)^2}{-\lambda_2}\right)} \quad (19)$$

Then by substituting Eq. (19) into Eq. (18), we can derive the solution of the alternative optimal weight for the t-th view as follows:

$$w_t = \frac{\exp\left(\frac{\sum_{i=1}^{N}\sum_{k=1}^{K} \widetilde{u}_{ik,t,\eta} \sum_{j\in G_t} m_{j,t}\left(x_{ij,t} - v_{kj,t}\right)^2}{-\lambda_2}\right)}{\sum_{t'=1}^{T} \exp\left(\frac{\sum_{i=1}^{N}\sum_{k=1}^{K} \widetilde{u}_{ik,t',\eta} \sum_{j\in G_{t'}} m_{j,t'}\left(x_{ij,t'} - v_{kj,t'}\right)^2}{-\lambda_2}\right)} \quad (20)$$

3.2.4 Partition matrix optimization

In this step, we update the partition matrix $u_{ik,t}$ by fixing variables V, W, and M in the t-th view.

With constraint $\sum_{k=1}^{K} u_{ik,t} = 1$ and through Lagrangian optimization, the minimization of $J_{\text{TW-Co-MFC}}(U, V, W, M)$ in Eq. (4) by fixing variables V, W and M is equivalent to the optimization of $u_{ik,t}$ as follows:

$$J_{\text{TW-Co-MFC}}(u_{ik,t}, \alpha_{i,t}) = \sum_{t=1}^{T} w_t \left(\sum_{i=1}^{N} \sum_{k=1}^{K} \Delta \sum_{j \in G_t} m_{j,t} \left(x_{ij,t} - v_{kj,t} \right)^2 \right) + \lambda_1 \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{k=1}^{K} u_{ik,t} \ln u_{ik,t} + \lambda_2 \sum_{t=1}^{T} w_t \ln w_t + \lambda_3 \sum_{t=1}^{T} \sum_{j \in G_t} m_{j,t} \ln m_{j,t} - \sum_{t=1}^{T} \sum_{i=1}^{N} \alpha_{i,t} \left(\sum_{k=1}^{K} u_{ik,t} - 1 \right) \tag{21}$$

where $\Delta = \left((1-\eta) u_{ik,t} + \dfrac{\eta}{T-1} \sum\limits_{t'=1, t' \neq t}^{T} u_{ik,t'} \right)$.

Then by setting $\partial J_{\text{TW-Co-MFC}} / \partial u_{ik,t} = 0$ and $\partial J_{\text{TW-Co-MFC}} / \partial \alpha_{i,t} = 0$, we obtain the membership degree of the i-th sample of the j-th cluster center in the t-th view as follows:

$$u_{ik,t} = \frac{\exp\left(\dfrac{w_t (1-\eta) d_{ik,t} + \sum\limits_{t'=1, t' \neq t}^{T} w_{t'} \dfrac{\eta}{T-1} d_{ik,t'}}{-\lambda_1} \right)}{\sum\limits_{k'=1}^{K} \exp\left(\dfrac{w_t (1-\eta) d_{ik',t} + \sum\limits_{t'=1, t' \neq t}^{T} w_{t'} \dfrac{\eta}{T-1} d_{ik',t'}}{-\lambda_1} \right)} \tag{22}$$

where $d_{ik,t} = \sum_{j \in G_t} m_{j,t} \left(x_{ij,t} - v_{kj,t} \right)^2$.

3.3 TW-Co-MFC algorithm

On the basis of the aforementioned inference, a TW-Co-MFC based on maximum entropy is designed. In each iteration, we update V, M, W, and U alternatively, and calculate the objective function value. We repeat this procedure until convergence or the number of iterations reaches the predefined maximum number of iterations. The procedure is described in detail in Algorithm 1.

The convergence of Algorithm 1 can be proven as follows. The clustering analysis task expressed in Eq. (4) is divided into four convex subproblems and each subproblem includes only one variable. After solving each subproblem alternatively to derive the optimal solutions, i.e., Eqs. (8), (14), (20), and (22), our algorithm can derive the optimal solution and converge to a local optimal solution.

3.4 Analysis of parameters λ_1, λ_2, and λ_3

Thus far, the fuzzy weighting strategy can be classified into two categories, namely, entropy weighting and fuzzy weighting methods[30]. In our proposed method, we not only adopt the entropy weighting approach to measure the importance of views and features but also use the maximum entropy principle to calculate the degree of the sample point belonging to a certain cluster. Given that the mathematical principle is the same, we only take the analysis of λ_1 as an example to show the influence of parameter variation on the fuzzy distribution.

The value range of parameter λ_1 is $(0, \infty)$. Then we rewrite Eq. (22) as follows:

Algorithm 1 TW-Co-MFC based on maximum entropy

Input:
The dataset $X = \{x_1, x_2, \ldots, x_N\}$ with N samples and T views, the three regularization parameters λ_1, λ_2, and λ_3, the number of clusters K, the trade-off factor η, the threshold ξ, and the maximum iteration time $\text{iter}_{\max}$.

Output:
The overall clustering membership matrix $\overline{U}$, the cluster centroid set V, the view weighting set W, and the feature weighting set M.

1: **Initialization**: Set $J_{\text{TW-Co-MFC}}(0) = 0$; randomly generate initialize assignment membership matrix U_t $\forall t$, initialize view weight as $W_t = 1/T$; initialize feature weight as $m_{j,t} = 1/|G_t|$ $\forall j \in G_t$; and set iter $= 1$.
2: **repeat**
3: Compute the cluster center V_t $\forall t$ via Eq. (8).
4: Update the feature weight M_t, $\forall t$ via Eq. (14).
5: Update the view weight W_t, $\forall t$ via Eq. (20).
6: Update the membership matrix U_t, $\forall t$ via Eq. (22).
7: iter $=$ iter $+ 1$.
8: **until** $|J_{\text{TW-Co-MFC}}(\text{iter}) - J_{\text{TW-Co-MFC}}(\text{iter}-1)| < \xi$ or iter $>$ $\text{iter}_{\max}$
9: Compute the overall membership matrix $\overline{U}$ via Eq. (6).
10: **return** The overall clustering membership matrix $\overline{U}$, the view weighting set W, and the feature weighting set M.

$$u_{ik,t}=\frac{\exp\left(\dfrac{D_{ik,t}}{-\lambda_1}\right)}{\sum\limits_{k'=1}^{K}\exp\left(\dfrac{D_{ik',t}}{-\lambda_1}\right)} \tag{23}$$

where $D_{ik,t} = w_t(1-\eta)\sum\limits_{j\in G_t} m_{j,t}\left(x_{ij,t}-v_{kj,t}\right)^2 + \sum\limits_{t'=1,t'\neq t}^{T} w_{t'}\dfrac{\eta}{T-1}\sum\limits_{j\in G_{t'}} m_{j,t'}\left(x_{ij,t'}-v_{kj,t'}\right)^2$.

Notably, $D_{ik,t}$ represents the weighted distance from sample i to cluster k in the t-th view. For any two randomly selected clusters l and s in the t-th view, $l \neq s$, from Eq. (23), we derive the following equation:

$$\frac{u_{il,t}}{u_{is,t}}=\frac{\exp\left(\dfrac{D_{il,t}}{-\lambda_1}\right)}{\exp\left(\dfrac{D_{is,t}}{-\lambda_1}\right)}=\exp\left(\frac{D_{il,t}-D_{is,t}}{-\lambda_1}\right) \tag{24}$$

From Eq. (24), we observe that the fuzzy membership distribution of the sample in the t-th view becomes uniform as the parameter λ_1 increases. In particular, when $\lambda_1 \to \infty$, we derive the following expression:

$$\lim_{\lambda_1\to\infty}\frac{u_{il,t}}{u_{is,t}}=\lim_{\lambda_1\to\infty}\frac{\exp\left(\dfrac{D_{il,t}}{-\lambda_1}\right)}{\exp\left(\dfrac{D_{is,t}}{-\lambda_1}\right)}= \lim_{\lambda_1\to\infty}\exp\left(\frac{D_{il,t}-D_{is,t}}{-\lambda_1}\right)=0 \tag{25}$$

In this case, the fuzzy membership of sample i in the t-th view will become uniform, which is not the result that we are expecting. When $\lambda_1 \to 0$, the fuzzy membership of sample i in the t-th view will become sharp, which is also not the desired result because it will lose the advantage of fuzzy partition.

Therefore, from the aforementioned analysis, we observe that the values of the three parameters can neither be too large nor too small.

3.5 Computational complexity analysis

In this subsection, we analyze the time complexity of the proposed algorithm. In each iteration, the sequential structure of TW-Co-MFC is divided into four key steps, which consist of the calculation of the cluster center, feature weight, view weight, and membership matrix for each view. For a dataset with N objects, D features, K clusters, and T views, the time complexity of each step is $O(NKD)$, which has nothing to do with the number of views in dataset. Assuming that TW-Co-MFC needs P iterations to converge, the overall time complexity is $O(PNKD)$.

4 Experimental Result and Discussion

In this section, extensive experiments are conducted on five benchmark datasets to evaluate the performance of the proposed algorithm. Firstly, TW-Co-MFC is compared with five existing baseline algorithms. Then the parameter impact analysis is conducted, i.e., the distribution of view weight W and clustering performance with different λ_2 values, the feature weight distribution M and clustering performance with different λ_3 values, and the influence of η on the clustering result. All of the experiments are conducted in MATLAB 2016a on a PC with 2 Intel 1.6 GHz CPU, 12 GB RAM, and Windows 10 64 bit.

4.1 Dataset

The experiments are performed on five real-world multi-view datasets. The statistics of these datasets are shown in Table 1. We briefly introduce the five datasets as follows:

- Handwritten (HW) is a handwritten digit dataset from the UCI machine learning repository[31]. The dataset contains 2000 samples classified into 10 classes. Each sample is one of the handwritten digits (0–9) described by 649 features divided into six views, namely, profile correlations (HW-fac), Fourier coefficients of the character shapes (HW-fou), Karhunen-loève coefficients (HW-kar), morphological features (HW-mor), pixel averages in 2 × 3 windows (HW-pix), and Zernike moments (HW-zer).
- Image Segmentation (IS) is a collection of outdoor images available at the UCI machine learning repository[31]. The dataset includes 2310 samples classified into 7 classes. Each sample is described by 19 features categorized into two views, i.e., shape and RGB (Red, Green, and Blue) views.
- Handwritten 2 sources (HW2sources) is a dataset consisting of 2000 samples collected from two sources[12], i.e., MNIST Handwritten Digits (0–9) and USPS Handwritten Digits (0–9).
- BBCSport contains 544 sport news articles collected from the BBC Sports website[32]. Each article

Table 1 Summary of the benchmark datasets.

Dataset	Number of objects	Number of views	Number of clusters	View structure
HW	2000	6	10	216-76-64-6-240-47
IS	2310	2	7	9-10
HW2sources	2000	2	10	784-256
BBCSport	544	2	5	3183-3203
Newsgroups	500	3	5	2000-2000-2000

was split into two segments (two views) and manually assigned to one of the five topical labels.

- Newsgroups consists of 500 documents obtained from the well-known 20-Newsgroup dataset[33]. Each raw document is classified into one of five topical labels after being preprocessed with three different feature extraction methods (considered as three views).

4.2 Evaluation measure

To evaluate the clustering performance, three widely used clustering performance measures are adopted on the basis of the ground-truth labels of the instances, which are Clustering Accuracy (ACC)[34], Normalized Mutual Information (NMI)[35], and Rand Index (RI)[36]. Note that the values of ACC, NMI, and RI vary from 0 to 1, with a higher value corresponding to a better clustering performance.

4.3 Baseline algorithm and other setting

The proposed algorithm is compared with three single-view and three multi-view clustering methods on the five real-world datasets mentioned in the previous subsection. To obtain comparable inertia for all views, we normalize all of the values of each feature in the dataset within the range $[-1, 1]$ before we perform any clustering. We select the standard FCM algorithm as the single-view counterpart of our method. The fuzzy degree parameter is set in the interval [1.05,1.5] with the step of 0.05, and the three other parameter values are adopted as the default settings. Firstly, we employ the FCM code available in the tool-box of MATLAB on every single view of the datasets and record the worst and best clustering results among different views. Then we apply the FCM to the concatenated features of all views. Co-FKM[28] is a fuzzy centralized method for multi-view clustering, which enables the collaboration between views following the disagreement-based strategy by generalizing the three fusion strategies, i.e., before, during or after the clustering process. MVKKM[37] is a kernel-based weighted multi-view clustering algorithm, in which each view is expressed in terms of given kernel matrices and a weighted combination of the kernels is learned in parallel to partitioning. TW-Co-k-means[15] is a two-level weighted collaborative k-means algorithm for multi-view clustering analysis, which works in a collaborative manner to take advantage of the complementary and consistent information across different views while simultaneously computing the weights for views and features. In our proposed method, the grid search strategy is adopted to identify the best parameters within the range of the candidate parameters, i.e., we take turns to fix three of the four parameters to obtain the optimal values and gradually vary the fourth parameter until TW-Co-MFC achieves the optima by grid search. The setting of the candidate parameters in these algorithms is given in Table 2. For all of the other approaches involved, the grid search strategy is also adopted to identify the optimal core parameters. To guarantee the reliability of the experimental results, we run each approach 10 times using random initialization and report the best average scores because all of the clustering algorithms depend on initialization. The optimal parameters of these methods that correspond

Table 2 Optimal parameter values for different algorithms on different datasets.

Algorithm	Candidate parameter	Optimal parameter				
		HW	IS	HW2sources	BBCSport	Newsgroups
Worst single view	m : [1.05 : 0.05 : 1.50]	1.50	1.10	1.35	1.05	1.20
Best single view	m : [1.05 : 0.05 : 1.50]	1.05	1.30	1.05	1.40	1.05
Concatenated view	m : [1.05 : 0.05 : 1.50]	1.25	1.05	1.10	1.35	1.45
Co-FKM	m : [1.10 : 0.05 : 1.50]	1.25	1.35	1.05	1.45	1.45
	η : [0.1 : 0.05 : $(T-1)/T$]	0.15	0.10	0.30	0.30	0.60
MVKKM	$p = [1, 1.3, 1.5, 2, 4, 6]$	6.0	2.0	1.3	1.3	1.3
TW-Co-k-means	α : [10 : 10 : 100]	45	70	40	50	30
	β : [1 : 10]	10	10	8	1	9
	η : [0.1 : 0.05 : 1]	0.45	0.45	0.35	0.20	0.20
TW-Co-MFC	x : [−9 : 0.5 : 3]					
	λ_1 : $\exp(x)$	exp(−8)	exp(−6)	exp(−4.5)	exp(−8.5)	exp(−8)
	λ_2 : $\exp(x)$	240	340	620	2	exp(5)
	λ_3 : $\exp(x)$	exp(7)	390	7500	60	exp(8)
	η : [0.1 : 0.05 : $(T-1)/T$]	0.25	0.50	0.40	0.50	0.40

to the best results are outlined in Table 2.

4.4 Comparison result

In this section, we evaluate the performance of our approach against five baseline approaches in terms of three clustering measurements, i.e., ACC, NMI, and RI on five benchmark datasets. Tables 3–5 show the comparison results where the values in bold indicate the best performance results among the seven algorithms.

From Tables 3–5, we observe that concatenating the features of all views directly to clustering analysis cannot guarantee that better clustering results than the best single-view results can be generated. For instance, the result of the concatenated IS dataset is worse than its best single-view result. The reason for this phenomenon can be the different discriminatory capabilities of different features and views. TW-Co-MFC is based on the two-level weight mechanism, which helps explore the different partitioning capabilities of views and features for the data containing compatible or incompatible views. In addition, TW-Co-MFC improves Co-FKM by adding two kinds of weights. We observe that TW-Co-MFC exhibits a better performance than the concatenating method and Co-FKM and even outperforms the TW-Co-k-means, which is a crisp two-level weight collaborative multi-view clustering method.

4.5 Parameter analysis

In this subsection, we conduct the parameter analysis of the proposed method, which consists of the view weight parameter λ_2, the feature weight parameter λ_3, and the collaborative factor η. In the process of analyzing one of the four parameters, all other parameters are fixed. Given the limitation in terms of the length of the paper, we only report the results of the HW and IS datasets. Table 6 summarizes the characteristics of these two datasets in detail.

4.5.1 View weight parameter λ_2

To investigate the influence of parameter λ_2 on the view

Table 3 Clustering performance comparison of each algorithm in terms of ACC on five real-world datasets.

Algorithm	ACC				
	HW	IS	HW2sources	BBCSport	Newsgroups
Worst single view	0.4577	0.5958	0.1860	0.3311	0.3180
Best single view	0.5704	0.6432	0.4533	0.3719	0.3280
Concatenated view	0.9080	0.6395	0.8920	0.5070	0.3632
Co-FKM	0.9116	0.6692	0.8970	0.4906	0.3834
MVKKM	0.8750	0.5706	**0.9320**	0.3621	0.2100
TW-Co-k-means	0.8063	0.6222	0.4987	0.4542	0.2904
TW-Co-MFC	**0.9214**	**0.6859**	0.8880	**0.5915**	**0.5120**

Table 4 Clustering performance comparison of each algorithm in terms of NMI on five real-world datasets.

Algorithm	NMI				
	HW	IS	HW2 sources	BBCSport	Newsgroups
Worst single view	0.4528	0.6175	0.1442	0.0290	0.1323
Best single view	0.5396	0.6410	0.4458	0.1229	0.1455
Concatenated view	0.8267	0.6093	0.8078	0.2454	0.2163
Co-FKM	0.8542	0.6268	0.8195	0.2516	0.2277
MVKKM	0.8116	0.6246	**0.8716**	0.0244	0.0230
TW-Co-k-means	0.8501	0.6180	0.4814	0.1857	0.1176
TW-Co-MFC	**0.8954**	**0.6715**	0.8302	**0.4352**	**0.4186**

Table 5 Clustering performance comparison of each algorithm in terms of RI on five real-world datasets.

Algorithm	RI				
	HW	IS	HW2sources	BBCSport	Newsgroups
Worst single view	0.8606	0.8600	0.5351	0.5094	0.4723
Best single view	0.8936	0.8769	0.8690	0.6089	0.5205
Concatenated view	0.9658	0.8730	0.9602	0.6811	0.5761
Co-FKM	0.9665	0.8778	0.9619	0.6761	0.5851
MVKKM	0.9578	0.8594	**0.9745**	0.2530	0.2129
TW-Co-k-means	0.9602	0.8662	0.8732	0.4587	0.3815
TW-Co-MFC	**0.9767**	**0.8879**	0.9641	**0.7385**	**0.6588**

Table 6 Detailed information of the HW and IS datasets.

Dataset	View	Number of features	Number of clusters	Number of instances
HW	HW-fac	216	10	2000
	HW-fou	76	10	2000
	HW-kar	64	10	2000
	HW-mor	6	10	2000
	HW-pix	240	10	2000
	HW-zer	47	10	2000
IS	Shape	9	7	2310
	RGB	10	7	2310

weight vector distribution and clustering performance, we set the range of the parameter λ_2 as $[60, 340]$ with intervals of 40 by fixing the other parameters as the optimal values listed in Table 2.

Figure 1 shows the distribution of the view weight as a function of λ_2 on the HW and IS datasets, with λ_2 ranging from 60 to 340. From Fig. 1, we observe that the smaller value of λ_2 leads to the sparser distribution of the view weight. With the increase in the value of λ_2, the distribution of view weight becomes more even, which indicates that the view weights become more uniform. In real-world application, if we have some prior knowledge about the usefulness of the data views, then we can select a relatively appropriate value of λ_2. Figure 2 shows the effect of parameter λ_2 in the range of $[60, 340]$ with the interval of 40 on the clustering performance in terms of

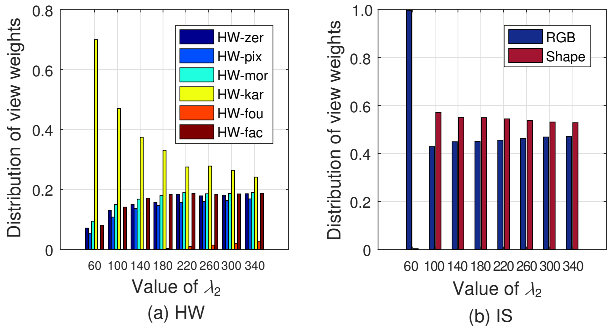

Fig. 1 Influence of λ_2 on distribution of view weights.

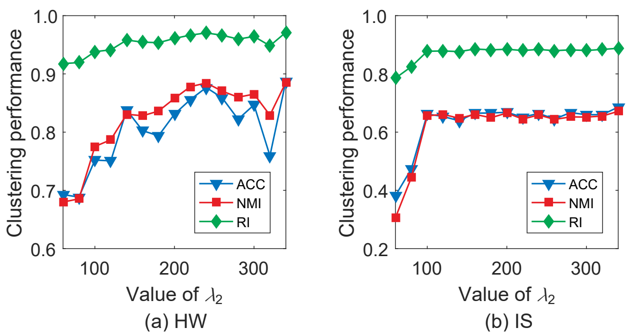

Fig. 2 Influence of λ_2 on clustering performance.

ACC, NMI, and RI. From Fig. 2, we observe that the clustering performance will be improved dramatically with the increase in the value of λ_2. However, when the parameter λ_2 exceeds a certain value, the clustering performance will decrease. For the HW dataset, the value is approximately 220. By contrast, for IS dataset, the value is approximately 140. The reason for this phenomenon is that, when λ_2 is set to have a small value, only one view is selected, resulting in the sparsest view selection and the loss of some useful view information. In another extreme case, when λ_2 is set to have a large value, the clustering performance will decrease because the view information with more noises will take part in the clustering assignment. Therefore, in real-world applications, we need to use a small λ_2 value for the dataset with incompatible views and a large λ_2 value for the dataset with compatible views.

4.5.2 Feature weight parameter λ_3

Following the method used to conduct weight parameter analysis, we set the range of parameter λ_3 as $\exp(x)$, where $x \in \{4.5, 5.5, 6.5, 7.5, 8.5\}$ for the HW dataset and $x \in \{2, 3, 4, 5, 6, 7\}$ for IS dataset by fixing the other parameters to have the optimal values listed in Table 2 to investigate the influence of λ_3 on the distribution of feature weight and clustering performance.

Figures 3 and 4 illustrate that the smaller the value of λ_3, the sparser the distribution of feature weights. Meanwhile, the larger the value of λ_3, the more uniform the distribution of feature weights. In practical applications, we need to set λ_3 to have appropriate values to balance the distribution of features on the basis of the characteristic of the features.

Figure 5 depicts the effect of parameter λ_3 on the clustering performance on the HW and IS datasets in terms of ACC, NMI, and RI. As shown in Fig. 5, on the two datasets, the clustering performance improves more dramatically when varying λ_3 from the defined smallest value to a certain value, e.g., exp(5) for the HW dataset and approximately exp(4) for the IS dataset. When exceeding the set value, compared with the situation when varying the parameter λ_2, the clustering results relatively fluctuate. The clustering performance is good when the parameter λ_3 is set in the range of exp(5)–exp(8.5) for the HW dataset and exp(4)–exp(6.5) for the IS dataset. However, when the parameter λ_3 is set to have a large value, the clustering performance shows a downward trend.

Furthermore, by combining the results of Figs. 3–5, we observe that a small λ_3 value is beneficial to the

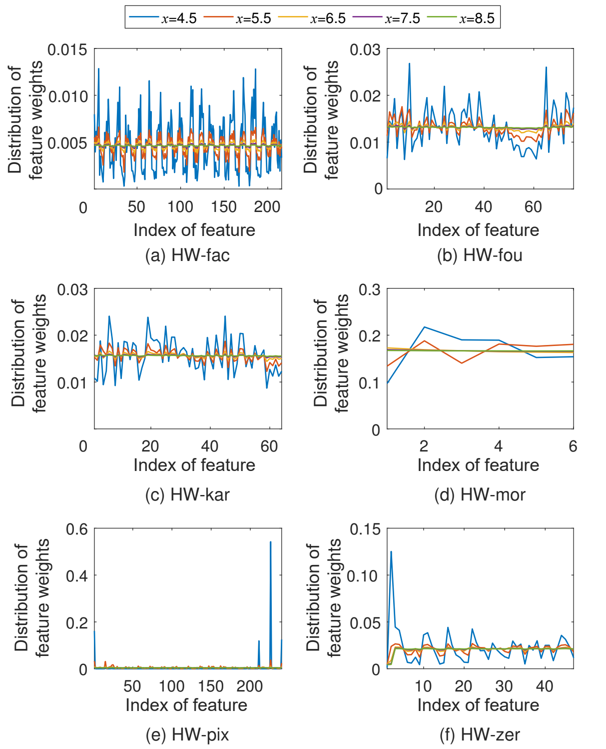

Fig. 3 Influence of λ_3 on distribution of feature weights on HW dataset.

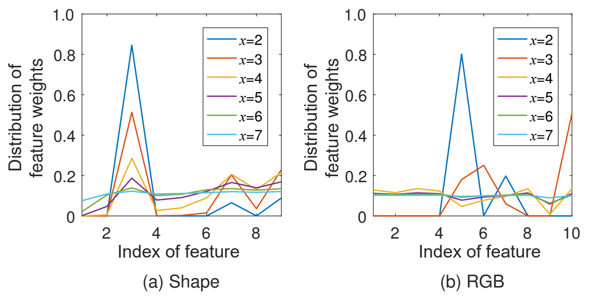

Fig. 4 Influence of λ_3 on distribution of feature weights on IS dataset.

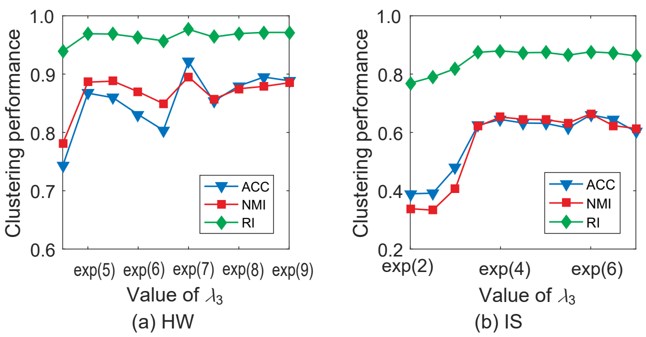

Fig. 5 Influence of λ_3 on clustering performance.

dataset with noise features, making the distribution of feature weights sharper. Meanwhile, a large λ_3 value makes the distribution of feature weights more equal so that more features contribute to the clustering. Therefore, on the basis of the comparison of the results of view weight analysis, we suggest that extreme discrimination of views and moderate discrimination of features can help achieve a good clustering performance.

4.5.3 Trade-off factor η

The trade-off factor η is used as a penalty associated with the disagreement between the membership matrices of different views, which would also affect the performance of our method. By fixing the three other parameters to have the optimal values listed in Table 2, we evaluate the performance of our algorithm with the trade-off factor η varying in the range of $\eta = [0.1 : 0.05 : (6-1)/6]$ for the HW dataset and $\eta = [0.1 : 0.05 : (2-1)/2]$ for the IS dataset. The results are displayed in Fig. 6. When the trade-off factor η tends to 0, the second term of Eq. (5) will tend to 0 and the algorithm will lose the view collaborative mechanism to fuse assignments during the clustering procedure. In another extreme situation, when the parameter equals $(T-1)/T$, all of the other views will have the same clustering assignment work to the objective function. Moreover, poor clustering assignment will occur during the calculation of the objective function, which will decrease the clustering performance. Hence, the trade-off factor η should have a moderate value, e.g., in the interval $[0.3, 0.6]$, which will help to obtain the most satisfactory clustering performance.

5 Conclusion

In this study, we propose a two-level collaborative multi-view soft clustering method based on maximum entropy to address the issue of uncertain clustering analysis. The proposed method has the following advantages compared with conventional multi-view clustering methods: (1) An adaptive two-level weighting process is designed to emphasize the importance of views and features within the same view simultaneously to express the inherently strong or weak discriminating

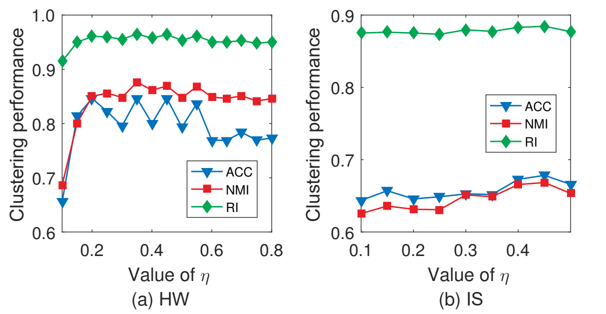

Fig. 6 Influence of η on clustering performance.

power of different views and features; (2) it utilizes a collaborative working mechanism to balance the quality of clusters in each view with the explicit clustering consistency between different views; and (3) a maximum-entropy based fuzzy multi-view clustering objective function is designed, which provides a better physical explanation for soft clustering partition than fuzzy c-means. Experiments on real-world multi-view datasets have demonstrated the effectiveness of our approach. In the future, we will consider the integration of rough set theory into the multi-view soft clustering process to effectively reduce the effect of outliers further.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Nos. 61603313, 61772435, 61976182, and 61876157).

References

[1] J. Zhao, X. J. Xie, X. Xu, and S. L. Sun, Multi-view learning overview: Recent progress and new challenges, *Inf. Fusion*, vol. 38, pp. 43–54, 2017.

[2] Y. Yang and H. Wang, Multi-view clustering: A survey, *Big Data Mining and Analytics*, vol. 1, no. 2, pp. 83–107, 2018.

[3] S. Bettoumi, C. Jlassi, and N. Arous, Collaborative multi-view K-means clustering, *Soft Comput.*, vol. 23, no. 3, pp. 937–945, 2019.

[4] J. Yu, Z. C. Qin, T. Wan, and X. Zhang, Feature integration analysis of bag-of-features model for image retrieval, *Neurocomputing*, vol. 120, pp. 355–364, 2013.

[5] Y. Z. Jiang, F. L. Chung, S. T. Wang, Z. H. Deng, J. Wang, and P. J. Qian, Collaborative fuzzy clustering from multiple weighted views, *IEEE Trans. Cybern.*, vol. 45, no. 4, pp. 688–701, 2015.

[6] B. Abu-Jamous, R. Fa, D. J. Roberts, and A. K. Nandi, Paradigm of tunable clustering using binarization of consensus partition matrices (Bi-CoPaM) for gene discovery, *PLoS One*, vol. 8, no. 2, p. e56432, 2013.

[7] X. Cai, F. P. Nie, and H. Huang, Multi-view K-means clustering on big data, in *Proc. 23rd Int. Joint Conf. Artificial Intelligence*, Beijing, China, 2013, pp. 2598–2604.

[8] X. J. Chen, X. F. Xu, J. Z. Huang, and Y. M. Ye, TW-k-means: Automated two-level variable weighting clustering algorithm for multiview data, *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 4, pp. 932–944, 2013.

[9] B. Jiang, F. Y. Qiu, and L. P. Wang, Multi-view clustering via simultaneous weighting on views and features, *Appl. Soft Comput.*, vol. 47, pp. 304–315, 2016.

[10] Y. M. Xu, C. D. Wang, and J. H. Lai, Weighted multi-view clustering with feature selection, *Pattern Recognit.*, vol. 53, pp. 25–35, 2016.

[11] S. D. Huang, Z. Kang, I. W. Tsang, and Z. L. Xu, Auto-weighted multi-view clustering via kernelized graph learning, *Pattern Recognit.*, vol. 88, pp. 174–184, 2019.

[12] H. Wang, Y. Yang, B. Liu, and H. Fujita, A study of graph-based system for multi-view clustering, *Knowl. Based Syst.*, vol. 163, pp. 1009–1019, 2019.

[13] W. Pedrycz, Collaborative fuzzy clustering, *Pattern Recognit. Lett.*, vol. 23, no. 14, pp. 1675–1686, 2002.

[14] C. D. Wang, J. H. Lai, and P. S. Yu, Multi-view clustering based on belief propagation, *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 4, pp. 1007–1021, 2016.

[15] G. Y. Zhang, C. D. Wang, D. Huang, W. S. Zheng, and Y. R. Zhou, TW-Co-k-means: Two-level weighted collaborative k-means for multi-view clustering, *Knowl. Based Syst.*, vol. 150, pp. 127–138, 2018.

[16] A. Cornuéjols, C. Wemmert, P. Gancarski, and Y. Bennani, Collaborative clustering: Why, when, what and how, *Inf. Fusion*, vol. 39, pp. 81–95, 2018.

[17] S. Zeng, X. Y. Wang, H. Cui, C. J. Zheng, and D. Feng, A unified collaborative multikernel fuzzy clustering for multiview data, *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 3, pp. 1671–1687, 2018.

[18] R. K. Xia, Y. Pan, L. Du, and J. Yin, Robust multi-view spectral clustering via low-rank and sparse decomposition. in *Proc. 28th AAAI Conf. Artificial Intelligence*, Québec City, Canada, 2014, pp. 2149–2155.

[19] F. P. Nie, J. Li, and X. L. Li, Self-weighted multiview clustering with multiple graphs, in *Proc. 26th Int. Joint Conf. Artificial Intelligence*, Melbourne, Australia, 2017, pp. 2564–2570.

[20] J. L. Xu, J. W. Han, F. P. Nie, and X. L. Li, Re-weighted discriminatively embedded K-means for multi-view clustering, *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 3016–3027, 2017.

[21] J. Hu, T. R. Li, C. Luo, H. Fujita, and Y. Yang, Incremental fuzzy cluster ensemble learning based on rough set theory, *Knowl. Based Syst.*, vol. 132, pp. 144–155, 2017.

[22] P. J. Qian, S. W. Sun, Y. Z. Jiang, K. H. Su, T. G. Ni, S. T. Wang, and R. F. Muzic Jr, Cross-domain, soft-partition clustering with diversity measure and knowledge reference, *Pattern Recognit.*, vol. 50, pp. 155–177, 2016.

[23] R. P. Li and M. Mukaidono, A maximum-entropy approach to fuzzy clustering, in *Proc. 1995 IEEE Int. Conf. Fuzzy Systems*, Yokohama, Japan, 1995, pp. 2227–2232.

[24] X. B. Zhi, J. L. Fan, and F. Zhao, Fuzzy linear discriminant analysis-guided maximum entropy fuzzy clustering algorithm, *Pattern Recognit.*, vol. 46, no. 6, pp. 1604–1615, 2013.

[25] P. J. Qian, Y. Z. Jiang, Z. H. Deng, L. Z. Hu, S. W. Sun, S. Wang, and R. F. Muzic, Cluster prototypes and fuzzy memberships jointly leveraged cross-domain maximum entropy clustering, *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 181–193, 2016.

[26] P. J. Qian, J. X. Zhou, Y. Z. Jiang, F. Liang, K. F. Zhao, S. T. Wang, K. H. Su, and R. F. Muzic, Multi-view maximum entropy clustering by jointly leveraging inter-view collaborations and intra-view-weighted attributes, *IEEE Access*, vol. 6, pp. 28 594–28 610, 2018.

[27] J. C. Bezdek, R. Ehrlich, and W. Full, FCM: The fuzzy c-means clustering algorithm, *Comput. Geosci.*, vol. 10, nos. 2&3, pp. 191–203, 1984.

[28] G. Cleuziou, M. Exbrayat, L. Martin, and J. H. Sublemontier, CoFKM: A centralized method for multiple-view clustering, in *Proc. 9th IEEE Int. Conf. Data Mining*, Miami, FL, USA, 2009, pp. 752–757.

[29] S. Miyamoto, H. Ichihashi, and K. Honda, *Algorithms for Fuzzy Clustering: Methods in C-Means Clustering with Applications*. Berlin, Germany: Springer, 2008.

[30] J. Wang, S. T. Wang, F. Chung, and Z. H. Deng, Fuzzy partition based soft subspace clustering and its applications in high dimensional data, *Inf. Sci.*, vol. 246, pp. 133–154, 2013.

[31] C. Blake and C. J. Merz, UCI repository of machine learning databases, http://archive.ics.uci.edu/ml/index.php, 1998.

[32] D. Greene and P. Cunningham, Practical solutions to the problem of diagonal dominance in kernel document clustering, in *Proc. 23rd Int. Conf. Machine Learning*, New York, NY, USA, 2006, pp. 377–384.

[33] S. F. Hussain, G. Bisson, and C. Grimal, An improved co-similarity measure for document clustering, in *Proc. 9th Int. Conf. Machine Learning and Applications*, Washington, DC, USA, 2010, pp. 190–197.

[34] J. Huang, F. P. Nie, H. Huang, and C. Ding, Robust manifold nonnegative matrix factorization, *ACM Trans. Knowl. Dis. Data*, vol. 8, no. 3, pp. 1–21, 2014.

[35] A. Strehl and J. Ghosh, Cluster ensembles—A knowledge reuse framework for combining multiple partitions, *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, 2003.

[36] L. Hubert and P. Arabie, Comparing partitions, *J. Classif.*, vol. 2, no. 1, pp. 193–218, 1985.

[37] G. Tzortzis and A. Likas, Kernel-based weighted multi-view clustering, in *Proc. IEEE 12th Int. Conf. Data Mining*, Brussels, Belgium, 2012, pp. 675–684.

Jie Hu received the BE degree from Southwest Normal University in 2001 and MS degree from East China Normal University in 2007. She received the PhD degree from Southwest Jiaotong University in 2016. In 2018–2019, she was a visiting scholar at the Georgia State University. She is currently an associate professor at School of Information Science and Technology, Southwest Jiaotong University. Her research interests include data mining, big data, clustering analysis, and clustering ensemble. She is the recipient of 2017 ACM Chengdu Doctoral Dissertation Award.

Yi Pan received the BE and ME degrees from Tsinghua University in 1982 and 1984, respectively, and the PhD degree from the University of Pittsburgh in 1991. He is currently a Regents' professor and chair of Computer Science Department at Georgia State University. He has served as an associate dean and chair of Biology Department during 2013–2017 and chair of Computer Science Department during 2006–2013. He joined Georgia State University in 2000, was promoted to full professor in 2004, named a distinguished university professor in 2013, and designated a Regents' professor (the highest recognition given to a faculty member by University System of Georgia) in 2015. His current research interests include parallel and cloud computing, big data, and bioinformatics. He has published more than 400 papers including over 230 SCI journal papers and 90 IEEE Transactions papers. In addition, he has edited/authored 43 books. His work has been cited more than 11 600 times based on Google Scholar and his current h-index is 56. He has served as an editor-in-chief or editorial board member for 20 journals including 7 IEEE Transactions. He is the recipient of many awards including one IEEE Transactions Best Paper Award, five IEEE and other international conference or journal Best Paper Awards, 4 IBM Faculty Awards, 2 JSPS Senior Invitation Fellowships, IEEE BIBE Outstanding Achievement Award, IEEE Outstanding Leadership Award, NSF Research Opportunity Award, and AFOSR Summer Faculty Research Fellowship. He has organized numerous international conferences and delivered keynote speeches at over 60 international conferences around the world.

Tianrui Li received the BS, MS, and PhD degrees from Southwest Jiaotong University in 1992, 1995, and 2002, respectively. He was a postdoctoral researcher at SCKCEN, Belgium from 2005 to 2006, and a visiting professor at Hasselt University, Belgium in 2008, University of Technology, Sydney, Australia in 2009, and University of Regina, Canada in 2014. He is currently a professor at School of Information Science and Technology, Southwest Jiaotong University and vice dean at Institute of Artificial Intelligence, Southwest Jiaotong University. He has authored or coauthored more than 300 research papers in refereed journals and conferences. His research interests include big data, cloud computing, data mining, granular computing, and rough sets. He is a fellow of IRSS and a senior member of ACM and IEEE.

Yan Yang received the BSc and MSc degrees from Huazhong University of Science and Technology in 1984 and 1987, respectively. She received the PhD degree from Southwest Jiaotong University in 2007. In 2002–2003 and 2004–2005, she was a visiting scholar at University of Waterloo. She is currently the professor and vice dean at School of Information Science and Technology, Southwest Jiaotong University. Her research interests include computational intelligence, big data, data mining, ensemble learning, and cloud computing. She has authored and coauthored over 100 research papers in refereed journals and conferences. She is vice chair of ACM Chengdu Chapter.