Collaborative Evidential Clustering

Yixuan Qiao, Shoumei Li and Thierry Denœux

Abstract Different companies may not be allowed to treat data together given restrictions of security, privacy or other technical reasons. In order to make better use of information from different sources, clustering algorithms based on collaboration mechanisms have been widely used. We propose the concept of collaborative evidential clustering under the framework of evidence theory. The key point is to establish collaboration among the credal partition matrices of each data site to meet the data confidentiality requirements. Considering the problems of excessive information interaction and insufficient information interaction, we design single-step and multi-step collaborative evidential clustering algorithms. Our algorithms were validated on real data sets.

1 Introduction

With the rapid development of information technology, the total amount of data is growing exponentially. Companies rely on their powerful storage capability to continuously collect, organize, and analyze data, in order to mine valuable information. A large amount of data is stored in different data sites and various types of servers. Due to security, privacy or other technical reasons, companies are reluctant to share data and only want to exchange information at non-data levels. In order to make bet-

Yixuan Qiao
Beijing University of Technology, 100 Pingleyuan, Chaoyang District Beijing, 100124, P.R. China, e-mail: yixuanqiao@bjut.edu.cn

Shoumei Li
Beijing University of Technology, 100 Pingleyuan, Chaoyang District Beijing, 100124, P.R. China, e-mail: lisma@bjut.edu.cn

Thierry Denœux
Université de Technologie de Compiègne, UMR CNRS 7253 Heudiasyc, France, e-mail: tdenoeux@utc.fr

ter use of different levels of information and reveal the internal information structure at local data sites, clustering algorithms based on collaborative mechanisms have been proposed. The basic idea of collaborative clustering is to first run a clustering algorithm independently at each data site, and then interact by exchanging the local structure information of each data site to reveal the potential common underlying structure of different data sites.

The concept of collaborative fuzzy clustering (CFC) and its implementation have been introduced in [15]. The development of CFC solves many practical problems (see, e.g., [12], [6], [5]). Collaboration mechanisms have been extensively studied from different perspectives: collaborative frameworks based on rough-fuzzy clustering or topological maps, higher-level collaborative schemes relying on existing clustering algorithms or collaborative approaches dedicated to distributed datasets (see, e.g., [14], [4], [7], [24]).

Since the original concept of CFC only implies a single collaboration phase, Pedrycz and Rai [16, 17] further refined the concept by making the collaboration an iterative process in which a specific data site would periodically use structure information from other data sites resulting from the collaboration process. In addition, experiments show that the performance of the CFC algorithm is not sensitive to the choice of the collaborative strength coefficient. In this paper, we study the specific form of the collaboration mechanism under the framework of Dempster-Shafer evidence theory.

Previous studies suggest that a direct comparison of two partition matrices after a few steps of collaboration could not be feasible as we may not have a direct correspondence between their rows (respective clusters). Consequently, several authors have proposed to rearrange the partition matrix (see e.g. [16], [17], [18]). Recently, the Hungarian algorithm [8] has been used to ensure that the same rows in the partition matrices refer to the same cluster. However, it has been found experimentally that the partition matrix reordering is not necessary, especially when facing a large number of phases of the collaboration [21].

Collaborative clustering has not been studied under the framework of evidence theory, which is considered to be a very mature theoretical system for uncertainty inference and widely used in many fields (see, e.g., [3], [9], [10], [2], [11]). Evidential clustering relies on the concept of credal partition [3], which uses mass functions to characterize the uncertainty in data effectively. The concept of credal partition extends those of hard, fuzzy and possibilistic partitions, and it constitutes a more general clustering framework. For each object, masses are assigned not only to single classes, but also to unions of classes. Experiments reported in [3] and [13] show that this extra flexibility allows us to have a deeper understanding of the data structure and improve the robustness to outliers.

In this paper, we study the implementation of a *collaborative evidential clustering* (CEC) algorithm, assuming the data at each site have the same sample size, the same number of clusters, and different feature spaces. Specifically, we explore the implementation of the collaboration mechanism in the ECM algorithm [13]. We consider a collaborative mechanism based on cluster structure information given existing data confidentiality requirements. Furthermore, to address the *Excessive In-*

formation Interaction (EII) and *Insufficient Information Interaction* (III) problems, we propose a single-step CEC algorithm and a multi-step CEC algorithm separately, in which the number of multi-step collaborations is controlled based on a structural similarity index (see e.g. [16], [17], [21]).

This paper is organized as follows. Section 2 recalls the background notions about belief functions, the ECM algorithm and the pignistic transform. The single-step and multi-step CEC algorithms are are introduced in Sections 3 and 4, respectively. Section 5 presents experimental results and some observations. Conclusions are given in Section 6.

2 Preliminaries

The Dempster-Shafer theory of evidence [20, 23] (or belief function theory) is a theoretical framework for representing partial and unreliable information. Let us consider a variable ω taking values in a finite set $\Omega = \{\omega_1, \ldots, \omega_j, \ldots, \omega_c\}$, called the frame of discernment. Partial knowledge regarding the actual value taken by ω can be represented by a *mass function* m, which is an application from the power set of Ω in the interval [0, 1] such that $\sum_{A \subseteq \Omega} m(A) = 1$. The subsets A of Ω such that $m(A) > 0$ are called the *focal sets* of m. The mass $m(A)$ can be interpreted as a fraction of a unit mass of belief that is allocated to A and that cannot be allocated to any subset of A. Complete ignorance is obtained when Ω is the only focal set, and full certainty when the whole mass of belief is assigned to a unique singleton of Ω. If all the focal sets of m are singletons, m is similar to a probability distribution: it is then called a *Bayesian* mass function. In the following, we use the concise notation $m_{ij} \in [0,1]$ to denote the belief of object x_i to subset A_j, and $m_{i\emptyset}$ to denote the mass assigned to the empty set. A mass function m_i such that $m_{i\emptyset} = 0$ is said to be *normalized*. Under the *open-world* assumption, the mass $m_{i\emptyset}$ is interpreted as a quantity of belief given to the hypothesis that the actual value of ω might not belong to Ω [22].

ECM is one of the algorithms proposed to derive a *credal partition* from data [13]. Deriving a credal partition implies determining, for each object x_i, the quantities $m_{ij} = m_i(A_j)$ in such a way that a low value of m_{ij} is found when the distance d_{ij} between x_i and A_j is high. In this framework, partial knowledge regarding the class membership of an object is represented by a mass function on the set of possible classes. Thus, belief mass may be given to any subset A of Ω (any set of classes), and not only to singletons of Ω. This representation makes it possible to model a wide variety of situations ranging from complete ignorance to full certainty. The ECM algorithm searches for the credal partition matrix M and cluster center matrix V that minimize the following criterion:

$$J_{ECM}(M,V) = \sum_{i=1}^{N} \sum_{\{j|A_j \neq \emptyset, A_j \subseteq \Omega\}} c_j^{\alpha} m_{ij}^2 d_{ij}^2 + \sum_{i=1}^{N} \delta^2 m_{i\emptyset}^2, \tag{1}$$

subject to the constraints $m_{ij} \geq 0$ for all i and j, $m_{i\emptyset} \geq 0$ for all i, and

$$\sum_{\{j|A_j\neq\emptyset,A_j\subseteq\Omega\}} m_{ij} + m_{i\emptyset} = 1, \tag{2}$$

for all i. In (1), $c_j = |A_j|$ is the cardinality of A_j and δ represents the distance of any object to the empty set.

In order to make a decision regarding the value of ω, it is possible to transform a normalized mass function m into a probability distribution using the following pignistic transformation [23]:

$$BelP(\omega) = \sum_{\{A\subseteq\Omega|\omega\in A\}} \frac{m(A)}{|A|}, \qquad \forall\omega\in\Omega. \tag{3}$$

3 A Single-Step CEC Algorithm

To meet data confidentiality requirements, we mainly consider the collaborative mechanism based on cluster structure information (including M and V) and integrate it into the objective function of the ECM algorithm.

The T data sites are denoted by $D[1],\ldots,D[t],\ldots,D[T]$. All data sites have the same number of samples denoted by N, same clusters denoted by c, but different feature spaces composed of $n[1],\ldots,n[t],\ldots,n[T]$ features, respectively. Matrix $K_{T\times T}$ is used to quantify the collaboration strength between each pair of data sites; its general term, denoted by $\kappa[t,s]$, represents the collaboration strength between sites $D[t]$ and $D[s]$. As there is no collaboration between $D[t]$ and itself, we set $\kappa[t,t]=0$.

For a given data site $D[t]$, we combine the information provided by the local data with the cluster structure information of the collaborators to determine the cluster structure. For that purpose, the objective function is expanded into the form

$$\begin{aligned} Q[t] = &\sum_{i=1}^{N}\sum_{\{j|A_j\neq\emptyset,A_j\subseteq\Omega\}} c_j^{\alpha}[t]m_{ij}^2[t]d_{ij}^2[t] + \sum_{i=1}^{N}\delta^2 m_{i\emptyset}^2[t] \\ &+ \sum_{s=1,s\neq t}^{T}\kappa[t,s]\sum_{i=1}^{N}\sum_{\{j|A_j\neq\emptyset,A_j\subseteq\Omega\}}(m_{ij}[t]-m_{ij}[s])^2 d_{ij}^2[t], \end{aligned} \tag{4}$$

for $t=1,2,\ldots,T$. The objective function $Q[t]$ contains two parts. The first part is the sum of weighted distances between the patterns in $D[t]$ and the center of the non-empty subset A_j; it is just the objective function (1) of the standard ECM applied to $D[t]$ with $\beta=2$. The second part implements the collaborative mechanism, which makes the clustering based on the $D[t]$ aware of other collaborators. The difference between $M[t]$ and $M[s]$ represents the difference in cluster structure between data sites $D[t]$ and $D[s]$. The weight $\kappa[t,s]$ controls the balance between local data information and collaborator cluster structure information. When $\kappa[t,s]=0$, the problem translates into a scenario where the standard ECM algorithm acts at each data site without collaboration. In general, we propose to constrain $\kappa[t,s]$ to be in the interval

$[0,1]$, so that the second term in the right-hand side of (4) does not dominate the first one.

As in the ECM algorithm, we also need $M[t]$ to satisfy constraints (2). Therefore, the single-step collaborative clustering algorithm consists in minimizing $Q[t]$ in (4) subject to (2). This optimization task splits into two problems, namely, determining the credal partition matrix $M[t]$ and the cluster center matrix $V[t]$. To determine the partition matrix, we exploit the technique of Lagrange multipliers. This leads to the new objective function that is formed separately for each data site $D[t]$, namely,

$$\begin{aligned} L[t] = &\sum_{i=1}^{N} \sum_{\{j|A_j\neq\emptyset, A_j\subseteq\Omega\}} c_j^{\alpha}[t] m_{ij}^2[t] d_{ij}^2[t] + \sum_{i=1}^{N} \delta^2 m_{i\emptyset}^2[t] \\ &+ \sum_{s=1,s\neq t}^{T} \kappa[t,s] \sum_{i=1}^{N} \sum_{\{j|A_j\neq\emptyset, A_j\subseteq\Omega\}} (m_{ij}[t] - m_{ij}[s])^2 d_{ij}^2[t] \\ &- \sum_{i=1}^{N} \lambda_i \left(\sum_{\{j|A_j\neq\emptyset, A_j\subseteq\Omega\}} m_{ij}[t] + m_{i\emptyset}[t] - 1 \right), \end{aligned} \tag{5}$$

where λ_i denotes a Lagrange multiplier. The necessary conditions leading to the local minimum of $M[t]$ read as follows:

$$\frac{\partial L[t]}{\partial m_{ij}[t]} = 2c_j^{\alpha}[t] m_{ij}[t] d_{ij}^2[t] + 2 \sum_{s=1,s\neq t}^{T} \kappa[t,s](m_{ij}[t] - m_{ij}[s]) d_{ij}^2[t] - \lambda_i = 0, \tag{6a}$$

$$\frac{\partial L[t]}{\partial m_{i\emptyset}[t]} = 2\delta^2 m_{i\emptyset}[t] - \lambda_i = 0, \tag{6b}$$

$$\frac{\partial L[t]}{\lambda_i} = \sum_{\{j|A_j\neq\emptyset, A_j\subseteq\Omega\}} m_{ij}[t] + m_{i\emptyset}[t] - 1 = 0. \tag{6c}$$

Introducing the notations

$$\psi[t] = \sum_{s=1,s\neq t}^{T} \kappa[t,s], \quad \varphi_{ij}[t] = \sum_{s=1,s\neq t}^{T} \kappa[t,s] m_{ij}[s], \tag{7}$$

we get the solution

$$m_{ij}[t] = \frac{\varphi_{ij}[t]}{c_j^{\alpha}[t] + \psi[t]} + \frac{\frac{1}{d_{ij}^2[t](c_j^{\alpha}[t]+\psi[t])} \left(1 - \sum_{\{j|A_j\neq\emptyset, A_j\subseteq\Omega\}} \frac{\varphi_{ij}[t]}{c_j^{\alpha}[t]+\psi[t]} \right)}{\sum_{\{j|A_j\neq\emptyset, A_j\subseteq\Omega\}} \frac{1}{d_{ij}^2[t](c_j^{\alpha}[t]+\psi[t])} + \frac{1}{\delta^2}}, \tag{8a}$$

$$m_{i\emptyset}[t] = 1 - \sum_{\{j|A_j\neq\emptyset, A_j\subseteq\Omega\}} m_{ij}[t]. \tag{8b}$$

In the calculations of the prototypes we confine ourselves to the weighted Euclidean distance between the sample and the centroid of the cluster, so the necessary

condition for solving the local minimum of the cluster center $V[t]$ is

$$\frac{\partial L[t]}{\partial v_l[t]} = \sum_{i=1}^{N} \sum_{\{j|A_j\neq\emptyset, A_j\subseteq\Omega\}} c_j^{\alpha}[t] m_{ij}^2[t] \frac{\partial d_{ij}^2[t]}{\partial v_l[t]} + \sum_{s=1,s\neq t}^{T} \kappa[t,s] \sum_{i=1}^{N} \sum_{\{j|A_j\neq\emptyset, A_j\subseteq\Omega\}} (m_{ij}[t] - m_{ij}[s])^2 \frac{\partial d_{ij}^2[t]}{\partial v_l[t]}. \quad (9)$$

Introducing the notations

$$B_{lq}[t] = \sum_{i=1}^{N} x_{iq} \sum_{w_l\in A_j} c_j^{\alpha-1} m_{ij}^2[t], \quad B_{lq}[t,s] = \sum_{i=1}^{N} x_{iq} \sum_{w_l\in A_j} (m_{ij}[t] - m_{ij}[s])^2 \frac{1}{c_j[t]},$$

$$H_{lk}[t] = \sum_{i=1}^{N} \sum_{\{w_k,w_l\}\subseteq A_j} c_j^{\alpha-2} m_{ij}^2[t], \quad H_{lk}[t,s] = \sum_{i=1}^{N} \sum_{\{w_k,w_l\}\subseteq A_j} (m_{ij}[t] - m_{ij}[s])^2 \frac{1}{c_j^2},$$

the cluster center matrix $V[t]$ has the form

$$V[t] = \left(H[t] + \sum_{s=1,s\neq t} \kappa[t,s] H[t,s] \right)^{-1} \left(B[t] + \sum_{s=1,s\neq t} \kappa[t,s] B[t,s] \right). \quad (10)$$

More details about the derivation process can be found in [19]. The algorithm can be described in Algorithm 1. It is worth noting that for each data site, the information used by single-step collaboration comes from the cluster structure information obtained by the initial ECM algorithm, not from the improved cluster structure information of the collaborator data site through collaboration.

Termination criterion I relies on the changes to the cluster center matrices obtained in successive iterations of the single-step CEC algorithm; we chose the L_∞ norm as a measure of change in the cluster center matrices. Subsequently, the optimization is terminated when this distance is lower than an assumed threshold value $\varepsilon > 0$.

4 A Multi-step CEC Algorithm

As the single-step CEC algorithm described in Section 3 may face the EII and III problems, we consider a multi-step collaboration mechanism to get more information from the collaborator data site for better interaction. The original purpose of collaboration was to reconcile and optimize the differences between cluster structures of various data sites. As the reconciliation continues, we can expect that the cluster structure similarity between the data sites will gradually increase. Therefore, we can use the structural similarity index to guide the multi-step CEC algorithm.

Algorithm 1 Single-step CEC Algorithm.

Require: $D[1],\ldots,D[t],\ldots,D[T]$, c, termination criterion I, $\kappa[t,s]$, ε
for $t=1$ **to** T **do**
  Use ECM to get the original $M_{original}[t]$ and $V_{original}[t]$
  $M_{final}[t] \leftarrow M_{original}[t]$, $V_{final}[t] \leftarrow V_{original}[t]$
end for
for $t=1$ **to** T **do**
  $l \leftarrow 0$, $I^0[t] \leftarrow 1$
  $M^0[t] \leftarrow M_{original}[t]$ and $V^0[t] \leftarrow V_{original}[t]$
  while $I^l[t] \geq \varepsilon$ **do**
    for $q=1$ **to** T **do**
      if $q \neq t$ **then**
        $M^l[q] \leftarrow M_{original}[q]$ and $V^l[q] \leftarrow V_{original}[q]$
      end if
    end for
    $l \leftarrow l+1$
    Compute $M^l[t]$, $V^l[t]$ using (8) and (10) with $M^{l-1}[t]$, $V^{l-1}[t]$
    $I^l[t] \leftarrow \max_{k\in[1,n[t]],\ j\in\{j|A_j\neq\phi,A_j\subseteq\Omega\}}(|V^l_{jk}[t]-V^{l-1}_{jk}[t]|)$
  end while
  $M_{final}[t] \leftarrow M^l[t]$, $V_{final}[t] \leftarrow V^l[t]$
end for
return $M_{final}[t]$ and $V_{final}[t]$

We should stress the fact that a direct comparison of two credal partition matrices could not be feasible as we may not have a direct correspondence between their rows (respective clusters). In a more general setting, we might even have different numbers of clusters at the individual data sites, and this diversity could make any attempt to form the correspondence between the partition matrices infeasible. Instead, we consider the following approach in which we test how the structure revealed at one data site performs on the remaining ones. Let us consider the following local structural similarity index:

$$W[t]=\sum_{s=1,s\neq t}^{T}\sum_{i=1}^{N}\sum_{\{j|A_j\neq\emptyset,A_j\subseteq\Omega\}} m_{ij}^2[s]\|x_i[t]-v_j[t|s]\|^2. \tag{11}$$

where

$$v_j[t|s]=\frac{\sum_{i=1}^{N} m_{ij}^2[s]x_i[t]}{\sum_{i=1}^{N} m_{ij}^2[s]}. \tag{12}$$

The rationale behind this measure is that if the structure of $D[s]$ is similar to that of $D[t]$, then the structure should also obtain a good performance on $D[t]$ (a more similar structure should lead to a lower value of $W[s]$). Finally, for all data sites, we have the following global structural similarity metric

$$W=\sum_{t=1}^{T} W[t]. \tag{13}$$

This indicator is used to control the number of iterations of the multi-step CEC algorithm, denoted as termination criterion II. So when $|W^l - W^{l-1}| < \varepsilon$, the multi-step CEC algorithm is completed.

The multi-step collaborative evidential clustering algorithm consists of three phases. It can be described in Algorithm 2.

Algorithm 2 Multi-step CEC Algorithm.

Require: $D[1],\ldots,D[t],\ldots,D[T]$, c, termination criterion I, termination criterion II, $\kappa[t,s]$, ε
1: Run Steps (1)-(20) of Algorithm 1, return values are denoted as $M^0[t]$ and $V^0[t]$
2: $l' \leftarrow 0$, II $\leftarrow W^0$
3: **while** II $\geq \varepsilon$ **do**
4: **for** $t = 1$ **to** T **do**
5: $M_{original}[t] \leftarrow M^{l'}[t]$, $V_{original}[t] \leftarrow V^{l'}[t]$
6: **end for**
7: $l' \leftarrow l' + 1$
8: Run (5)-(20) of Algorithm 1, return values are denoted as $M^{l'}[t]$ and $V^{l'}[t]$
9: $II \leftarrow |W^{l'} - W^{l'-1}|$
10: **end while**
11: **return** $M^{l'}[t]$ and $V^{l'}[t]$

The multi-step collaboration process is actually a cascade of single-step collaboration processes. We can imagine that before each single-step collaboration, the structural information of all data sites enters the information interaction pool, representing all the information that can be used in this single-step CEC process. For each data site, the collaborator cluster structure information that can be utilized is constant, only its own structure is constantly changing.

The final partition is determined by assigning each object to the cluster after convergence of the algorithm with maximal pignistic probability (3). Based on the given reference partition, we use the adjusted Rand index (ARI) to characterize the local collaboration quality of each data site. The global collaborative quality assessment index is then defined as

$$AARI = \frac{1}{T}\sum_{ii=1}^{T} ARI[ii]. \tag{14}$$

5 Experimental Results

In this section, we report on experimental findings[1] for some machine learning data sets [1]. The intent is to demonstrate the effectiveness of the collaboration and get some experimental insights into the behavior of the algorithms.

[1] We also conducted a simulation study. Experiments with synthetic dataset can be found in [19]. The ARI index for each data site is close to 1, which demonstrates that our algorithms are very competitive. Due to space limitations, the results of simulation studies have to be omitted.

The details of the features contained in the data site in each dataset are as follows. For Iris dataset: (a) Sepal.Length, Sepal.Width; (b) Sepal.Width, Petal.Length; (c) Petal.Length, Petal.Width. For Seeds dataset: (a) area, perimeter, compactness; (b) compactness, length of kernel, width of kernel; (c) width of kernel, asymmetry coefficient, length of kernel groove. We set $c = 3$, $\kappa[t,s] = 1$, $\varepsilon = 0.0001$. The evolutions of W and $AARI$ as a function of the number of iterations are reported in Figs. 1 and 2.

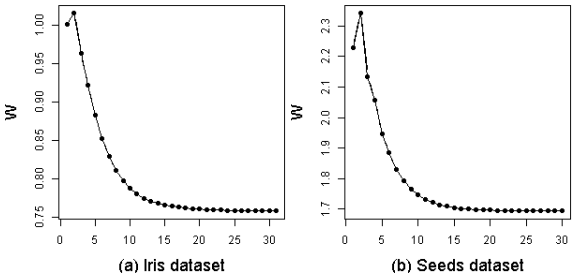

Fig. 1 With the continuous reconciliation of multi-step collaboration, the structural similarity between data sites is enhanced, and the value of W constantly declines. The EII problem emerged in the early stage, which resulted in a slight increase of W. The post-correction function of multi-step collaboration is indispensable.

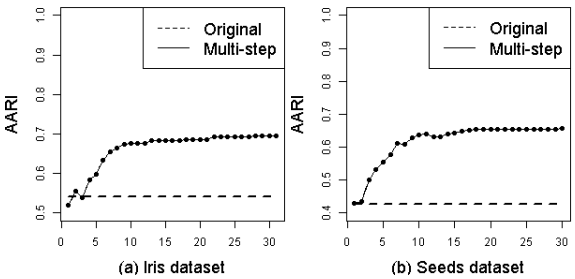

Fig. 2 EII and III led to a decrease in AARI after single step collaboration; the multi-step collaboration process performs effective information correction, making the index AARI strictly superior to the initial value after multi-step CEC. This further confirms the adequacy and necessity of the multi-step collaborative process.

At the local level, the structural similarity indicators of all data sites in each data set further confirm the existence of EII and III problems (see Fig. 3). Furthermore, we find that indices W and W[t] have roughly the same trend, which shows that W is excellent at global direction control.The ARI and its reference level for each data set are shown in Fig. 4.

For data sites D[1] and D[3] in the Iris dataset, we found that in the first few collaborations, the information was redistributed multiple times to find the correct direction of collaboration, which laid an important foundation for the subsequent collaboration. In the algorithm debugging phase, we find that the randomness introduced by redistribution is the key factor that determines the improvement of the final multi-step CEC algorithm. For data site D[3] in the Wine dataset, we found that ARI decreased slightly in the late stage but was still significantly higher than the baseline level, while the ARIs of data sites D[1] and D[2] were still rising. This further confirms that collaboration is a long term interaction process. The information data site D[3] lost can cause the data sites D[1] and D[2] to get a bigger boost.

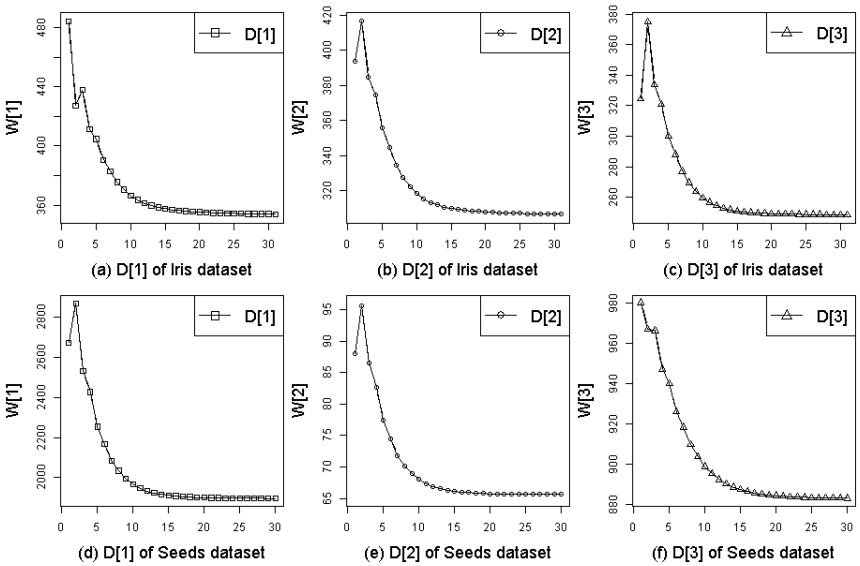

Fig. 3 The trend of local structural similarity index of Iris and Seeds dataset.

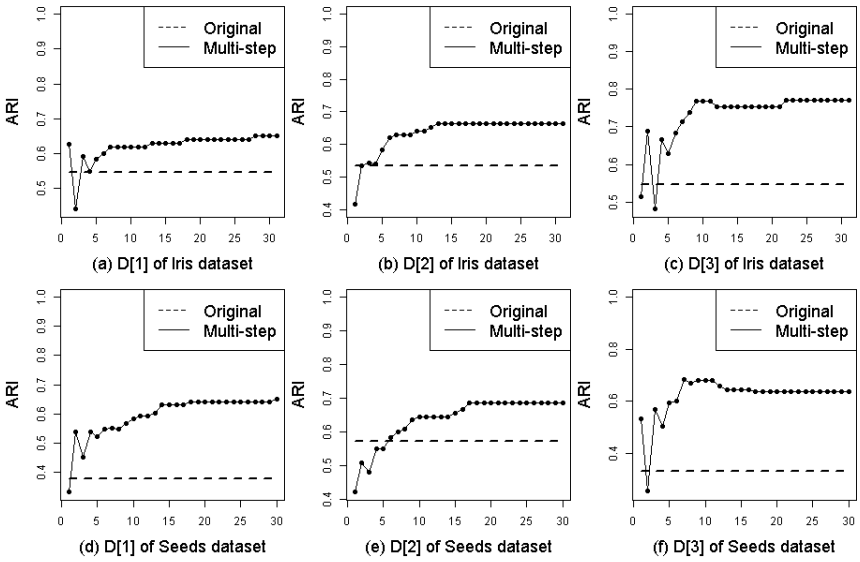

Fig. 4 Trend of ARI index of Iris and Seeds dataset.

Therefore, the loss is beneficial in general. It also indirectly shows that our choice of the global structural similarity index W as the stopping criterion of the multi-step CEC algorithm is reasonable. The numerical results of the multi-step CEC of each dataset are shown in Table 1.

Table 1 Multi-step collaborative clustering results analysis of Iris and Wine datasets

Data site	Original ARI	Single step ARI	Multi-step ARI	Promotion (%)
$Iris-D[1]$	0.546	0.626	0.652	0.106
$Iris-D[2]$	0.534	0.416	0.664	0.130
$Iris-D[3]$	0.547	0.515	0.771	0.224
$Wine-D[1]$	0.377	0.334	0.651	0.275
$Wine-D[2]$	0.571	0.423	0.687	0.115
$Wine-D[3]$	0.333	0.521	0.636	0.303

From Table 1, we can see more clearly that the single-step cooperation algorithm has poor stability and is prone to excessive information interaction, which causes the cluster structure to change too much and affect the value of ARI. The multi-step collaboration algorithm has a good correction effect, and it can improve ARI well after the information is redistributed. By further observation, we also found that the data site with a lower initial ARI value eventually rose significantly. The information it can use comes from the original data site with a higher value of ARI and a data site with a large difference from its cluster structure.

6 Concluding Remarks

In this study we have proposed a new concept of collaborative evidential clustering. Our multi-step CEC algorithm has been validated on real data sets and the experimental results have shown competitive performances. The EII and III problems in the single step CEC algorithm play a very good role in information redistribution and lay the foundation for multi-step CEC algorithm. So far, the research we have completed is based on two assumptions: (a) all data sites have the same number of clusters, (b) the strength of collaboration between different data sites is the same. These assumptions will be relaxed in future work.

Acknowledgements This work was supported by National Nature Science Foundation of China (No. 11571024), and by a grant to the third author as part of the Overseas Talent program from the Beijing Government. Shoumei Li (email: lisma@bjut.edu.cn) is the corresponding author of this paper.

References

1. Asuncion, A., Newman, D.: UCI machine learning repository (2007)

2. Bordes, J. B., Davoine, F., Xu, P., Denœux, T.: Evidential grammars: A compositional approach for scene understanding. Application to multimodal street data. Applied Soft Computing, **61**, 1173-1185 (2017)
3. Denœux, T., Masson, M. H.: EVCLUS: evidential clustering of proximity data. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), **34**(1), 95-109 (2004)
4. Depaire, B., Falcn, R., Vanhoof, K., Wets, G.: PSO driven collaborative clustering: A clustering algorithm for ubiquitous environments. Intelligent Data Analysis, **15**(1), 49-68 (2011)
5. Elhamifar, E., Vidal, R.: Sparse subspace clustering: Algorithm, theory, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, **35**(11), 2765-2781 (2013)
6. Forestier, G., Wemmert, C., Gancarski, P.: Collaborative multi-strategical clustering for object-oriented image analysis. In: Supervised and Unsupervised Ensemble Methods and their Applications, pp. 71-88. Springer, Berlin, Heidelberg (2008)
7. Ghassany, M., Grozavu, N., Bennani, Y.: Collaborative generative topographic mapping. Proceedings of International Conference on Neural Information Processing, pp. 591-598. Springer, Berlin, Heidelberg (2012)
8. Kuhn, H. W.: The Hungarian method for the assignment problem. Naval Research Logistics (NRL), **52**(1), 7-21 (2005)
9. Lelandais, B., Ruan, S., Denœux, T., Vera, P., Gardin, I.: Fusion of multi-tracer PET images for dose painting. Medical Image Analysis, **18**(7), 1247-1259 (2014)
10. Lian, C., Ruan, S., Denœux, T., Jardin, F., Vera, P.: Selecting radiomic features from FDG-PET images for cancer treatment outcome prediction. Medical Image Analysis, **32**, 257-268 (2016)
11. Lian, C., Ruan, S., Denœux, T., Li, H., Vera, P.: Spatial evidential clustering with adaptive distance metric for tumor segmentation in FDG-PET images. IEEE Transactions on Biomedical Engineering, **65**(1), 21-30 (2018)
12. Loia, V., Pedrycz, W., Senatore, S.: Semantic web content analysis: A study in proximity-based collaborative clustering. IEEE Transactions on Fuzzy Systems, **15**(6), 1294-1312 (2007)
13. Masson, M. H., Denœux, T.: ECM: An evidential version of the fuzzy c-means algorithm. Pattern Recognition, **41**(4), 1384-1397 (2008)
14. Mitra, S., Banka, H., Pedrycz, W.: Rough–fuzzy collaborative clustering. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), **36**(4), 795-805 (2006)
15. Pedrycz, W.: Collaborative fuzzy clustering. Pattern Recognition Letters, **23**(14), 1675-1686 (2002)
16. Pedrycz, W., Rai, P.: A multifaceted perspective at data analysis: a study in collaborative intelligent agents. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), **38**(4), 1062-1072 (2008)
17. Pedrycz, W., Rai, P.: Collaborative clustering with the use of Fuzzy C-Means and its quantification. Fuzzy Sets and Systems, **159**(18), 2399-2427 (2008)
18. Prasad, M., Siana, L., Li, D. L., Lin, C. T., Liu, Y. T., Saxena, A.: A preprocessed induced partition matrix based collaborative fuzzy clustering for data analysis. Proceedings of IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 14), pp. 1553-1558 (2014)
19. Qiao, Y.: On study of collaborative evidential clustering algorithm with applications. Dissertation for M.S. degree of Beijing University of Technology under the supervision of Li S. and Denoeux T (2019)
20. Shafer, G.: A mathematical theory of evidence. Princeton university press (1976)
21. Shen, Y., Pedrycz, W.: Collaborative fuzzy clustering algorithm: Some refinements. International Journal of Approximate Reasoning, **86**, 41-61 (2017)
22. Smets, P.: The transferable belief model for quantified belief representation. In: Quantified Representation of Uncertainty and Imprecision, pp. 267-301. Springer, Dordrecht (1998)
23. Smets, P., Kennes, R.: The transferable belief model. Artificial Intelligence, **66**(2), 191-234 (1994)
24. Sublime, J., Grozavu, N., Bennani, Y., Cornujols, A.: Collaborative clustering with heterogeneous algorithms. Proceedings of IEEE International Joint Conference on Neural Networks (IJCNN-IEEE 15), pp. 1-8 (2015)