Coupled Clustering Ensemble: Incorporating Coupling Relationships Both between Base Clusterings and Objects

Can Wang, Zhong She, Longbing Cao

Advanced Analytics Institute, University of Technology, Sydney, Australia
{canwang613,zhong2024,longbing.cao}@gmail.com

***Abstract*— Clustering ensemble is a powerful approach for improving the accuracy and stability of individual (base) clustering algorithms. Most of the existing clustering ensemble methods obtain the final solutions by assuming that base clusterings perform independently with one another and all objects are independent too. However, in real-world data sources, objects are more or less associated in terms of certain coupling relationships. Base clusterings trained on the source data are complementary to one another since each of them may only capture some specific rather than full picture of the data. In this paper, we discuss the problem of explicating the dependency between base clusterings and between objects in clustering ensembles, and propose a framework for coupled clustering ensembles (*CCE*). *CCE* not only considers but also integrates the coupling relationships between base clusterings and between objects. Specifically, we involve both the intra-coupling within one base clustering (i.e., cluster label frequency distribution) and the inter-coupling between different base clusterings (i.e., cluster label co-occurrence dependency). Furthermore, we engage both the intra-coupling between two objects in terms of the base clustering aggregation and the inter-coupling among other objects in terms of neighborhood relationship. This is the first work which explicitly addresses the dependency between base clusterings and between objects, verified by the application of such couplings in three types of consensus functions: clustering-based, object-based and cluster-based. Substantial experiments on synthetic and UCI data sets demonstrate that the *CCE* framework can effectively capture the interactions embedded in base clusterings and objects with higher clustering accuracy and stability compared to several state-of-the-art techniques, which is also supported by statistical analysis.**

I. Introduction

Clustering ensemble [1] has exhibited great potential in enhancing the clustering accuracy, robustness and parallelism [2] by combining results from various clustering methods. Its objective is to produce an overall high-quality clustering that agrees as much as possible with each of the input clusterings. Clustering ensemble can be applied in various settings, such as clustering heterogeneous data or privacy-preserving information [3]. In general, the whole process of clustering ensemble can be divided into three parts: building base clusterings, aggregating base clusterings, and post-processing clustering. While the clustering ensemble largely captures the common structure of the base clusterings, and achieves a combined clustering with better quality than that of individual clusterings, it also faces several issues that have not been explored well in the consensus design. We illustrate the problem with the related work and the challenge of clustering ensemble below.

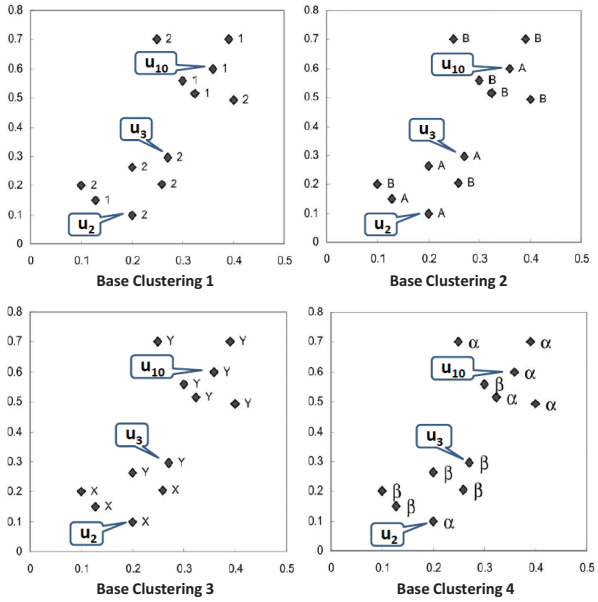

Fig. 1. Four possible base clusterings of 12 data objects into two clusters, different partitions use different sets of labels.

Taking the clustering ensemble described in Fig. 1 [4] as an example, it shows four two-cluster partitions of 12 two-dimensional data objects. The target of clustering ensemble is to obtain a final clustering based on these four base clusterings. As shown in Fig. 1, the four possible cluster labels for the objects u_2, u_3 and u_{10} are $\{2, A, X, \alpha\}$, $\{2, A, Y, \beta\}$ and $\{1, A, Y, \alpha\}$, respectively. That is to say, two of the four base clusterings put each pair of objects in the same group, and the rest two partitions assign different cluster labels to this pair. For instance, the first and second base clusterings distribute u_2 and u_3 in the same cluster, while the last two base clusterings give distinct labels to them. In this situation, the traditional clustering ensemble method (i.e., *CSPA* [2]) treats the similarity between every pair of these three objects to be 0.5, which is $Sim(u_2, u_3) = Sim(u_2, u_{10}) = Sim(u_3, u_{10}) = 0.5$. In the last stage of post-processing clustering, thus, it is difficult to determine the final label for these objects. The reason is that the similarity defined here is too limited to reveal

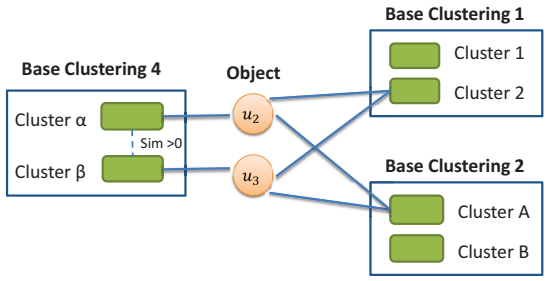

Fig. 2. A graphical representation of the coupled relationship between base clusterings, where each circle denotes an object, each rectangle represents an cluster, and an edge exists if an object belongs to a cluster.

the complete hidden relationship among the data set from the initial results of base clustering. A conventional way is to randomly distribute them in either an identical cluster or different groups, which will inevitably affect the clustering performance.

However, if we explore the information provided in Fig. 1 carefully, we are able to identify some coupling relationships between the base clusterings and between the data objects, apart from the consensus among initial results proposed by traditional ensemble strategies.

On one hand, as indicated in Fig. 2, objects u_2 and u_3 are considered to have a high similarity value (e.g., 1) in base clusterings 1 and 2, in which they are assigned to the same clusters (i.e., cluster 2 and cluster A, respectively). In contrast, their similarity value is rather low (e.g., 0) if only the information in base clustering 4 is used, since they are grouped into different clusters: α and β. However, cluster α and cluster β are intuitively more similar than they appear to be, due to the fact that they connect with two identical clusters in other base clusterings via objects u_2 and u_3. Thus, the similarity (i.e., the dashed line) between clusters α and β related with other base clusterings should be larger than 0. The same principle also applies to the similarity between clusters X and Y in base clustering 3. Note that here the similarity between the same clusters (e.g., cluster A or cluster 2) is manually set to be 1. In this way, the overall similarity between objects u_2 and u_3 must be larger than 0.5 as traditional method considers. Accordingly, objects u_2 and u_3 are more likely to be assigned to the same cluster as they should be, rather than depending on the random distribution.

On the other hand, the similarity between objects u_2 and u_3 and that between u_2 and u_{10} are identical (i.e., both 0.5). Thus, how to distinguish them and assign the correct label to each object? If we just consider the aforementioned coupled relationship between base clusterings, we may fail since both similarities will be enhanced by involving the co-occurrence with the clusters in other base clusterings. However, we discover that the discrepancy on the common neighborhood domains of objects u_2 with u_3 and u_2 with u_{10} is capable to differentiate u_2 and u_{10} in distinct clusters. Intuitively, we notice that in Fig. 1, the number of common neighbors of objects u_2 and u_3 is much larger than that of u_2 and u_{10}. From this perspective, it is more probable that objects u_2 and u_3 are in the same cluster and object u_{10} is in another one, which just corresponds to the genuine partition.

Based on the above issues, we then come up with three research questions in the following.

1) *Clustering Coupling*: There is likely structural relationship between base clusterings since they are induced from the same data set. How to describe the coupling relationship between base clusterings?
2) *Object Coupling*: There is context surrounding two objects which makes them dependent on each other. How to design the similarity or distance between objects to capture their relation with other data objects?
3) *Integrated Coupling*: If there are interactions between both clusterings and objects, then how to integrate such couplings in clustering ensemble?

Intuitively, the base clusterings are expected to have some interactions with each other, such as the co-occurrence of their cluster labels over the same set of objects. Here, the cluster label refers to the label of a cluster to which an object belongs, such as α, β in Fig. 1. But most of the existing methods, such as *CSPA* [2] and *QMI* [4], are all based on the hypothesis that base clusterings are independent of each other. Furthermore, the similarity between any two objects within the same cluster is not always the same and should be distinguished. In the existing work, however, the available approaches mostly treat the similarity between objects to be roughly 1 if they belong to the same cluster, otherwise 0. Such a binary measure is rather rough in terms of capturing the relationships between objects. In addition, some controversial objects with approximately equal similarity are observed to have different sizes of common neighborhood domain to differentiate them apart. But the current approaches have not addressed this issue for clustering ensemble problem, they merely consider the similarity between a pair of objects irrespective of other objects. Recently, a link-based approach [5] has been proposed to consider the cluster-cluster similarity by connected-triple approach, which shows promising progress. But it overlooks the interaction between objects. Besides, a clustering algorithm *ROCK* for categorical data has been introduced by Guha et al. [6] to specify the interaction of objects. However, it is just designed for the categorical clustering and lacks the consideration on the relationship between base clusterings. For *integrated coupling*, no work has been reported that systematically takes into account the couplings between base clusterings and between data objects.

In the real world, business and social applications such as investors in capital markets and members in social networking almost always see objects coupled with each other [7]. There is a great demand from both practical and theoretical perspectives to initiate new mechanisms to explicitly address the couplings both between base clusterings and between objects, and to explicate how to incorporate the couplings for clustering ensemble based on consensus functions.

In this paper, we propose an effective framework for coupled clustering ensembles (*CCE*) to address the aforementioned research questions. The key contributions are as follows:

– We consider both the couplings between base clusterings and between data objects, and propose a coupled framework of clustering ensembles to form an integrated coupling.
– We explicate our proposed framework *CCE* from the perspectives of clustering-based, object-based, and cluster-based algorithms, and reveal that the couplings are essential to clustering ensemble.
– We evaluate our proposed framework *CCE* with existing clustering ensemble and categorical clustering algorithms on a variety of benchmark data sets in terms of accuracy, stability, and statistical significance.

The paper is organized as follows. In Section II, we briefly review the related work. Preliminary definitions are specified in Section III. Section IV proposes the coupled framework *CCE*. Coupled relationships between base clusterings and between objects in *CCE* are specified in Section V. Section VI presents the coupled consensus functions for *CCE* together with miscellaneous issues. We describe the *CCE* algorithms in Section VII. The effectiveness of *CCE* is shown in Section VIII with extensive experiments. We conclude this work and address future work in Section IX.

II. Related Work

Several papers [2], [3] address the issue of consensus function for clustering ensemble. Heuristics include *CSPA*, *HGPA* and *MCLA* [2] solve the ensemble problem by firstly transforming the base clusterings into a hypergraph representation. Further, Fern and Brodley [8] further proposed *HBGF* to consider the similarities between objects and between clusters collectively. Gionis et al. [3] mapped the clustering aggregation problem to the weighted correlation clustering problem with linear cost functions. Besides, Topchy et al. [4] introduced a mixture probability model *EM* and an information-theoretic consensus function *QMI* to effectively combine weak base clusterings. Most of the existing research has been summarized in [9], in which the equivalence is revealed between basic partition difference (*PD*) algorithm and other advanced methods such as Chi-squared based approaches.

All the above methods, either fail to address the interactions between base clusterings and between objects (e.g., *CSPA*, *QMI*) or just assume the independence between them (e.g., *EM*). In particular, the weighted correlation clustering solution proposed in [3] fails to partition the objects if their distance measures are equally 0.5. However, an increasing number of researchers argue that clustering ensemble is also dependent on the relationship between input partitions [5], [10], [11]. Punera and Ghosh [10] put forward the soft cluster ensembles, in which they used a fuzzy clustering algorithm for the generation of base clusterings. The weighted distance measure [11] represents a soft relation between a pair of object and cluster. Unlike our proposed framework, those refined solutions of different base clusterings are stacked up to form the consensus function without explicitly addressing the relations among input clusterings. More recently, Iam-On et al. [5] presented a link-based approach to involve the cluster-cluster similarity based on the interaction between clusters. However, our method also squeezes out the intra-coupling within base clusterings and the relationship between objects, which means this work [5] forms just a part of our framework.

Alternatively, clustering ensemble can also be regarded as categorical clustering by treating each base clustering as an attribute [3]. We only illustrate the widely used categorical clustering algorithms here. Guha et al. [6] proposed *ROCK*, which uses the link-based similarity between two categorical objects. Andritsos et al. [12] introduced *LIMBO* that quantifies the relevant information preserved when clustering. In summary, *ROCK* considers the relationship between objects by link; *LIMBO* concerns the interaction between different attributes. Neither of them takes couplings between attributes and between objects into account together, whereas our proposed framework addresses both.

Besides, in our previous work [13], we proposed a coupled nominal similarity measure to specify the coupling of attributes. In this paper, we focus on a coupled framework for clustering ensemble, which addresses a new problem with different challenges and also involves the coupling of objects.

III. Preliminary Definitions

The problem of clustering ensemble can be formally described as follows: $U = \{u_1, \cdots, u_m\}$ is a set of m objects for clustering; $C = \{bc_1, \cdots, bc_L\}$ is a set of L base clusterings, each clustering bc_j consists of a set of clusters $bc_j = \{c_j^1, \cdots, c_j^{t_j}\}$ where t_j is the number of clusters in base clustering bc_j $(1 \leq j \leq L)$. Our goal is to find a final clustering $fc^* = \{c_*^1, \cdots, c_*^{t^*}\}$ with t^* clusters such that the objects inside each cluster c_*^t are close to each other and the objects in different clusters are far from each other.

We construct an information table S by mapping each base clustering as an attribute. Here, v_j^x indicates the label of a cluster to which the object u_x belongs in the jth base clustering, and V_j is the set of cluster labels in base clustering bc_j. For example, Table I [4] is the representation of Fig. 1 as an information table consisting of twelve objects $\{u_1, \cdots, u_{12}\}$ and four corresponding attributes (i.e., base clusterings $\{bc_1, bc_2, bc_3, bc_4\}$). The cluster label α in base clustering bc_4 is mapped as the attribute value v_4^2 of object u_2 on attribute bc_4, and cluster label set $V_4 = \{\alpha, \beta\}$.

Based on this information-table representation, we use several concepts adapted from our previous work [13]. The "set information function" $g_j(v_j^x)$ specifies the set of objects whose cluster labels is v_j^x in base clustering bc_j. For example, we have $g_4(v_4^2) = g_4(\alpha) = \{u_2, u_7, u_8, u_{10}, u_{11}, u_{12}\}$. We adopt the "inter-information function" $\varphi_{j \to k}(v_j^x)$ to obtain a subset of cluster labels in base clustering bc_k for the corresponding objects, which are derived from the cluster label v_j^x in base clustering bc_j, e.g., $\varphi_{4 \to 2}(\alpha) = \{A, B\}$ derived from object set $g_4(\alpha)$. Besides, the "information conditional probability" $P_{k|j}(v_k|v_j^x)$ characterizes the percentage of the objects whose cluster labels in base clustering bc_k is v_k among those objects whose cluster label in base clustering bc_j is exactly v_j^x,

TABLE I

AN EXAMPLE OF BASE CLUSTERINGS

U \ C	bc_1	bc_2	bc_3	bc_4
u_1	2	B	X	β
u_2	2	A	X	α
u_3	2	A	Y	β
u_4	2	B	X	β
u_5	1	A	X	β
u_6	2	A	Y	β
u_7	2	B	Y	α
u_8	1	B	Y	α
u_9	1	B	Y	β
u_{10}	1	A	Y	α
u_{11}	2	B	Y	α
u_{12}	1	B	Y	α

formalized as:

$$P_{k|j}(v_k|v_j^x) = \frac{|g_k(v_k) \cap g_j(v_j^x)|}{|g_j(v_j^x)|}, \tag{III.1}$$

where v_k is a fixed cluster label in base clustering bc_k. Note that $|\cdot|$ is the number of elements in the specific set. For example, we have $P_{2|4}(A|\alpha) = 2/6 = 1/3$.

All these concepts and functions form the foundation of the framework for capturing the coupled interactions between base clustering and between objects.

IV. COUPLED FRAMEWORK OF CLUSTERING ENSEMBLES

In this section, a coupled framework of clustering ensembles *CCE* is proposed in terms of both interactions between base clusterings and between data objects. In the framework described in Fig. 3, the coupling of base clusterings is revealed via the similarity between cluster labels v_j^x and v_j^y of each base clustering bc_j; and the coupling of objects is specified by defining the similarity between data objects u_x and u_y. In addition, three models are proposed for clustering-based, object-based, and cluster-based consensus building, revealing that the couplings are essential to clustering ensemble.

In terms of the *clustering coupling*, relationships within each base clustering and the interactions between distinct base clusterings are induced from the coupled nominal similarity measure *COS* in [13]. The intra-coupling of base clusterings captures the cluster label frequency distribution, while the inter-coupling of base clusterings considers the cluster label co-occurrence dependency [13]. *Object coupling* also focuses on the intra and inter-coupling, in which intra-coupling combines all the results of base clusterings for data objects, whereas inter-coupling is explicated by the neighborhood relationship [6] among different data objects. The object coupling also leads to a more accurate similarity ($\in [0, 1]$) between data objects. Moreover, as indicated in Fig. 3, the data objects and base clusterings are associated through the corresponding clusters, i.e., the position of an object in a clustering is determined by which cluster the object belongs to. Therefore, an integrated coupling is derived by treating each cluster label as an attribute value and then defining the similarity between objects grounded on the similarity between cluster labels over all base clusterings.

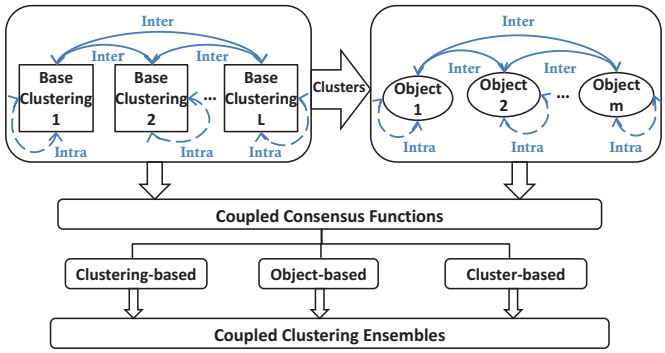

Fig. 3. A coupled framework of clustering ensembles (*CCE*), where ⇠----⇢ indicates the intra-coupling and ⟷ refers to the inter-coupling.

Given a set of m objects U and a set of L base clusterings C, we specify those interactions and the coupled consensus functions of *CCE* below in the following two sections.

V. COUPLED RELATIONSHIP IN *CCE*

In this section, we introduce how to describe the coupling of base clusterings and how to represent the coupling of objects.

A. Coupling of Clusterings

Since all the base clusterings are conducted on the same data objects, intuitively we assume there must be some relationship among those base clusterings. The coupling of base clusterings is proposed from the perspectives of intra-coupling and inter-coupling. The intra-coupling of base clusterings indicates the involvement of cluster label occurrence frequency within one base clustering, while inter-coupling of base clusterings means the interaction of other base clusterings with this base clustering [13]. Accordingly, we have:

Definition 5.1: (*IaCSC*) The **Intra-coupled Clustering Similarity for Clusters** between cluster labels v_j^x and v_j^y of base clustering bc_j is:

$$\delta_j^{IaC}(v_j^x, v_j^y) = \frac{|g_j(v_j^x)| \cdot |g_j(v_j^y)|}{|g_j(v_j^x)| + |g_j(v_j^y)| + |g_j(v_j^x)| \cdot |g_j(v_j^y)|}, \tag{V.1}$$

where $g_j(v_j^x)$ and $g_j(v_j^y)$ are the set information functions.

By taking into account the frequency of cluster labels, *IaCSC* characterizes the cluster similarity in terms of cluster label occurrence times. As clarified by [13], Equation (V.1) is a well-defined similarity measure and satisfies two main principles: greater similarity is assigned to the cluster label pair which owns approximately equal frequencies; the higher these frequencies are, the closer are the two clusters. For example, in Table I, we have $\delta_j^{IaC}(\alpha, \beta) = 3/4$.

IaCSC considers the interaction between cluster labels within an base clustering bc_j. It does not involve the coupling between base clusterings (e.g., between base clusterings bc_k and $bc_j (k \neq j)$) when calculating cluster label similarity. For this, we discuss the dependency aggregation, i.e., inter-coupled interaction.

Definition 5.2: (*IeRSC*) The **Inter-coupled Relative Similarity for Clusters** between cluster labels v_j^x and v_j^y of base clustering bc_j based on another base clustering bc_k is:

$$\delta_{j|k}(v_j^x, v_j^y|V_k) = \sum_{v_k \in \cap} \min\{P_{k|j}(v_k|v_j^x), P_{k|j}(v_k|v_j^y)\}, \tag{V.2}$$

where $v_k \in \cap$ denotes $v_k \in \varphi_{j\to k}(v_j^x) \cap \varphi_{j\to k}(v_j^y)$, $\varphi_{j\to k}$ is the inter-information function, and $P_{k|j}$ is the information conditional probability formalized in Equation (III.1).

Definition 5.3: (*IeCSC*) The **Inter-coupled Clustering Similarity for Clusters** between cluster labels v_j^x and v_j^y of base clustering bc_j is:

$$\delta_j^{IeC}(v_j^x, v_j^y|\{V_k\}_{k\neq j}) = \sum_{k=1,k\neq j}^{L} \lambda_k \delta_{j|k}(v_j^x, v_j^y|V_k), \tag{V.3}$$

where λ_k is the weight for base clustering bc_k, $\sum_{k=1,k\neq j}^{L} \lambda_k = 1$, $\lambda_k \in [0,1]$, $V_k (k \neq j)$ is a cluster label set of base clustering bc_k different from bc_j to enable the inter-coupled interaction, and $\delta_{j|k}(v_j^x, v_j^y|V_k)$ is *IeRSC*.

According to [13], relative similarity $\delta_{j|k}$ is an improved similarity measure derived from *MVDM* proposed by Cost and Salzberg [14]. It considers the similarity of two cluster labels v_j^x and v_j^y in base clustering bc_j on each possible cluster label in base clustering bc_k to capture the co-occurrence comparison between them. Further, the similarity δ_j^{IeC} between the cluster pair (v_j^x, v_j^y) in base clustering bc_j can be calculated on top of $\delta_{j|k}$ by aggregating all the relative similarity on base clusterings other than bc_j. For the parameter λ_k, in this paper, we simply assign $\lambda_k = 1/(L-1)$. For example, in Table I, we obtain $\delta_{4|2}(\alpha, \beta|V_2) = 1/3 + 1/2 = 5/6$ and $\delta_4^{IeC}(\alpha, \beta|\{V_1, V_2, V_3\}) = 1/3\times5/6+1/3\times5/6+1/3\times4/6 = 7/9$ if we take $\lambda_1 = \lambda_2 = \lambda_3 = 1/3$.

Thus, *IaCSC* captures the base clustering frequency distribution by calculating occurrence times of cluster labels within one base clustering, and *IeCSC* characterizes the base clustering dependency aggregation by comparing co-occurrence of the cluster labels in objects among different base clusterings. Finally, there is an eligible way to incorporate these two couplings together, specifically:

Definition 5.4: (*CCSC*) The **Coupled Clustering Similarity for Clusters** between cluster labels v_j^x and v_j^y of clustering bc_j is:

$$\delta_j^C(v_j^x, v_j^y|\{V_k\}_{k=1}^L) = \delta_j^{IaC}(v_j^x, v_j^y) \cdot \delta_j^{IeC}(v_j^x, v_j^y|\{V_k\}_{k\neq j}), \tag{V.4}$$

where δ_j^{IaC} and δ_j^{IeC} are *IaCSC* and *IeCSC*, respectively.

As indicated in Equation (V.4), *CCSC* gets larger by increasing either *IaCSC* or *IeCSC*. For example, in Table I, we could consider the coupled similarity of cluster labels α and β to be $\delta_j^C(\alpha, \beta|\{V_1, V_2, V_3, V_4\}) = 3/4 \times 7/9 = 7/12$.

Here, we choose the multiplication of these two components. The rationale is twofold: (1) *IaCSC* is associated with how often the cluster label occurs while *IeCSC* reflects the extent of the cluster difference brought by other base clusterings. Hence intuitively, the multiplication of them indicates the total amount of the cluster difference; (2) the multiplication method is consistent with the adapted simple matching distance introduced in [15], which considers both the category frequency and matching distance.

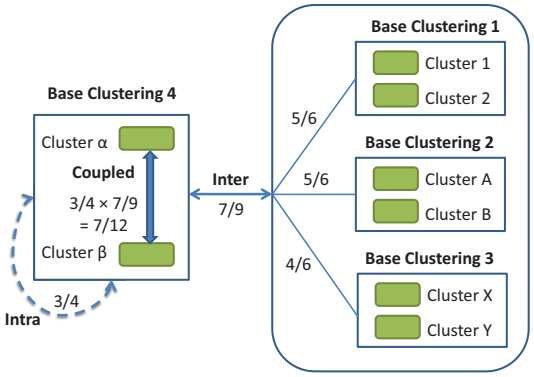

Fig. 4. An example of the coupled similarity for cluster labels α and β, where $\leftarrow\!\!\!\!\rightarrow$ indicates the intra-coupling and $\longleftrightarrow$ refers to the inter-coupling, the value along each line is the corresponding similarity.

Fig. 4 summarizes the whole process to calculate the coupled similarity for two cluster labels α and β. As indicated here, the similarity value between cluster labels α and β is $7/12$, which is larger than 0 as the existing methods regard. Thus, *CCSC* discloses the implicit relationship on both the frequency of cluster labels (intra-coupling) in each base clustering and the co-occurrence of cluster labels (inter-coupling) across different base clusterings.

B. Coupling of Objects

In the previous section, we present the coupling of base clusterings from the aspects of intra-coupled similarity and inter-coupled similarity between cluster labels. Here, we proceed by considering the coupling relationships among objects. Similarly, we assume the objects interact with each other both internally and externally.

In terms of the intra-perspective, the object u_x is coupled with u_y by involving the cluster labels of all the base clusterings for them. The similarity between u_x and u_y could be defined as the average sum of the similarity between the associated cluster labels ranging over all the base clusterings. Formally, we have:

Definition 5.5: (*IaOSO*) The **Intra-coupled Object Similarity for Objects** between objects u_x and u_y with respect to all the base clustering results of these two objects is:

$$\delta^{IaO}(u_x, u_y) = \frac{1}{L} \cdot \sum_{j=1}^{L} \delta_j^C(v_j^x, v_j^y|\{V_k\}_{k=1}^L), \tag{V.5}$$

where $\delta_j^C(v_j^x, v_j^y, \{V_k\}_{k=1}^L)$ refers to *CCSC* between cluster labels v_j^x and v_j^y of base clustering bc_j.

In this way, all the *CCSCs* δ_j^C $(1 \leq j \leq L)$ with each base clustering bc_j are summed up for two objects u_x and u_y. For example, the similarity between u_2 and u_3 in Table I is $\delta^{IaO}(u_2, u_3) = 0.655$ and $\delta^{IaO}(u_2, u_{10}) = 0.684$, which are both larger than 0.5 as provided by the traditional approach. We find that the intra-coupled object similarity between objects u_2 and u_{10} is a little bit greater than that between u_2 and u_3, which may somewhat mislead the final

clustering in the post-procession stage. To solve this problem, we also involve the coupling between objects to further expose the interaction on the object level.

As indicated in [6], the set theory-based similarity measure for categorical values, such as the Jaccard coefficient [15], often fails to capture the genuine relationship when the hidden clusters are not well-separated and there is a wide variance in the sizes of clusters. This is also true for our proposed *IaOSO*, since it considers the similarity between only the two objects in question as well. However, it does not reflect the properties of the neighborhood of the objects. Therefore, we present our new coupled similarity for objects based on the notions of neighbor and *IeOSO* as follows.

Definition 5.6: A pair of objects u_x and u_y are defined to be **neighbors** if the following holds:

$$\delta^{Sim}(u_x, u_y) \geq \theta, \tag{V.6}$$

where δ^{Sim} denotes any similarity measure for objects, $\theta \in [0, 1]$ is a given threshold.

In the above definition on neighbor, the similarity measure can be the Jaccard coefficient [6] for objects described by categorical attributes, or Euclidean dissimilarity [15] for objects depicted by continuous attributes. The neighbor set of object u_x can be denoted as:

$$N_{u_x}^{Sim} = \{u_z | \delta^{Sim}(u_x, u_z) \geq \theta\}, \tag{V.7}$$

which collects all the neighbors of u_x to form an object set N_{u_x}. For example, u_3 and u_{10} are the neighbors of object u_2, since $\delta^{Sim}(u_2, u_3) = \delta^{Sim}(u_2, u_{10}) = 1/3 \geq 0.3$ if we adopt the Jaccard coefficient as the similarity measure and set $\theta = 0.3$, and then the neighbor set of u_2 is $N_{u_2} = \{u_1, u_3, u_4, u_5, u_6, u_7, u_{10}, u_{11}\}$.

Further, we can embody the inter-coupled interaction between different objects by exploring the relationship between their neighborhood. Intuitively, objects u_x and u_y more likely belong to the same cluster if they have a larger overlapping in their neighbor sets N_{u_x} and N_{u_y}. Accordingly, below we use the common neighbors to define the inter-coupled similarity for objects.

Definition 5.7: (*IeOSO*) The **Inter-coupled Object Similarity for Objects** between objects u_x and u_y in terms of other objects u_z is defined as the ratio of common neighbors of u_x and u_y upon all the objects in U.

$$\delta^{IeO}(u_x, u_y | U) = \frac{1}{m} \cdot |\{u_z \in U | u_z \in N_{u_x}^{Sim} \cap N_{u_y}^{Sim}\}|, \tag{V.8}$$

where $N_{u_x}^{Sim}$ and $N_{u_y}^{Sim}$ are the neighbor sets of objects u_x and u_y based on δ^{Sim}, respectively.

Thus, *IeOSO* builds the inter-coupled relationship between each pair of objects by capturing the global knowledge on the neighborhood of them. For example, $\delta^{IeO}(u_2, u_3|U) = 0.583$ and $\delta^{IeO}(u_2, u_{10}|U) = 0.417$ when setting δ^{Sim} to be Jaccard coefficient and $\theta = 0.3$.

Finally, the intra-coupled and inter-coupled interactions could be considered together to induce the following coupled similarity for objects by exactly specializing the similarity measure δ^{Sim} in (V.7) to be *IaOSO* δ^{IaO} in Equation (V.5).

TABLE II
AN EXAMPLE OF NEIGHBORHOOD DOMAIN FOR OBJECT

Object	Neighborhood Domain
u_2	$\{u_1, u_3, u_4, u_5, u_6, u_7, u_8, u_{10}, u_{11}, u_{12}\}$
u_3	$\{u_1, u_2, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}, u_{11}, u_{12}\}$
u_{10}	$\{u_2, u_3, u_6, u_7, u_8, u_9, u_{11}, u_{12}\}$
Object Pair	**Common Neighbors**
u_2, u_3	$\{u_1, u_4, u_5, u_6, u_7, u_8, u_{10}, u_{11}, u_{12}\}$
u_2, u_{10}	$\{u_3, u_6, u_7, u_8, u_{11}, u_{12}\}$

Definition 5.8: (*CCOSO*) The **Coupled Clustering and Object Similarity for Objects** between objects u_x and u_y is defined when δ^{Sim} is in particular regarded as δ^{IaO}. Specifically:

$$\delta^{CO}(u_x, u_y | U) = \frac{1}{m} \cdot |\{u_z \in U | u_z \in N_{u_x}^{IaO} \cap N_{u_y}^{IaO}\}|, \tag{V.9}$$

where sets of objects $N_{u_x}^{IaO} = \{u_z | \delta^{IaO}(u_x, u_z) \geq \theta\}$ and $N_{u_y}^{IaO} = \{u_z | \delta^{IaO}(u_y, u_z) \geq \theta\}$.

In this way, the coupled similarity takes into account both the intra-coupled and inter-coupled relationships between two objects. At the same time, it also considers both the intra-coupled and inter-coupled interactions between base clusterings, since one of the components *IaOSO* of *CCOSO* is built on top of them. Thus, we call it the coupled clustering and object similarity for objects (*CCOSO*). For example, the corresponding neighbors of objects u_2, u_3 and u_{10} are described in the Table II below, here $\theta = 0.65$.

From this table, we observe that the number of common neighbors of objects u_2 and u_3 (i.e., 9) is truly larger than that of objects u_2 and u_{10} (i.e., 7), which correctly corresponds to our claim in Section I. Based on Equation (V.9), we obtain $\delta^{CO}(u_2, u_3|U) = 0.75$ and $\delta^{CO}(u_2, u_{10}|U) = 0.5$. It means that the similarity between objects u_2 and u_3 is larger than that between u_2 and u_{10}, which effectively remedies the issue caused by $\delta^{IaO}(u_2, u_3) < \delta^{IaO}(u_2, u_{10})$.

VI. COUPLED CONSENSUS FUNCTION IN *CCE*

There are many ways to define the consensus function such as pairwise agreements between base clusterings, co-associations between data objects, and interactions between clusters. Some of the criteria focus on the estimation of similarity between base clusterings [9], [4], some are based on the similarity between data objects [2], and others are associated with the similarity between clusters [8], [5]. In the following, we specify the coupled versions of clustering-based, object-based, and cluster-based criteria individually.

A. Clustering-based Coupling

The clustering-based consensus function captures the pairwise agreement between base clusterings. Note that each base clustering bc_j defines an associated similarity matrix $(BC_j)_{m \times m}$ that stores the information for each pair of objects about their similarity. Each entry $BC_j(x, y)$ of the matrix represents the similarity between objects u_x and u_y within the base clustering bc_j.

The usual way to define the entry $BC_j(x,y)$ of similarity matrix BC_j is to justify whether objects u_x and u_y are in the same cluster of base clustering bc_j, i.e., whether u_x and u_y have the same cluster label. Formally:

$$BC_j(x,y) = \begin{cases} 1 & \text{if } v_j^x = v_j^y, \\ 0 & \text{otherwise,} \end{cases} \tag{VI.1}$$

where v_j^x and v_j^y are the cluster labels of u_x and u_y in base clustering bc_j, respectively. Then, given two base clusterings bc_{j_1} and bc_{j_2}, a common measure of discrepancy is the partition difference (*PD*) [9]:

$$S_{Cg}(bc_{j_1}, bc_{j_2}) = \sum_{1 \le x,y \le m} \left[BC_{j_1}(x,y) - BC_{j_2}(x,y)\right]^2, \tag{VI.2}$$

where x and y refer to the indexes of objects u_x and u_y respectively. However, this traditional way is too rough to characterize the similarity between objects, and it assumes the independence among the base clusterings.

Alternatively, we can focus on the entry $BC_j(x,y)$ to incorporate the coupling of base clusterings as follows:

$$BC_j^C(x,y) = \delta_j^C(v_j^x, v_j^y | \{V_k\}_{k=1}^L), \tag{VI.3}$$

$$S_{Cg}^C(bc_{j_1}, bc_{j_2}) = \sum_{1 \le x,y \le m} \left[BC_{j_1}^C(x,y) - BC_{j_2}^C(x,y)\right]^2, \tag{VI.4}$$

where δ_j^C refers to *CCSC* in Definition 5.4. We denote this newly proposed clustering-based coupling to be *CgC*.

Intuitively, S_{Cg}^C calculates the sum of similarity between objects that belong to different base clusterings bc_{j_1} and bc_{j_2}. A target clustering fc^* thus should be:

$$fc^* = \underset{c^1,\cdots,c^{t^*}}{\arg} \min \sum_{j=1}^{L} S_{Cg}^C(fc, bc_j), \tag{VI.5}$$

where $fc = \{c^1, \cdots, c^{t^*}\}$ denotes the candidate set of clusters for final clustering fc^*. According to [4], the optimization problem in (VI.5) then can be heuristically approached by *k-means* operating in the normalized object-label space OL with each entry to be:

$$OL(u_x, v_j^y) = \delta_j^C(v_j^x, v_j^y | \{V_k\}_{k=1}^L) - \mu^y(\delta_j^C), \tag{VI.6}$$

where u_x is an object, v_j^y is a cluster label in bc_j, and $\mu^y(\delta_j^C)$ is the mean of $\delta_j^C(v_j^x, v_j^y | \{V_k\}_{k=1}^L)$ for every cluster.

B. Object-based Coupling

The object-based consensus function captures the co-associations between objects. Given two objects u_x and u_y, based on all the base clustering results, a simple and obvious heuristic to describe the similarity between u_x and u_y is the entry-wise average of the L associated similarity matrices induced by the L base clusterings. In this way, it yields an overall similarity matrix BC^* with a finer resolution [2]. Formally, we have:

$$BC^*(x,y) = \frac{1}{L} \cdot \sum_{j=1}^{L} BC_j(x,y). \tag{VI.7}$$

The entry of the induced overall similarity matrix BC^* is the weighted average sum of each associated pairwise similarity BC_j between objects of every base clustering. However, the common pairwise similarity measure $BC_j(x,y)$ is rather rough since only 1 and 0 are considered as defined in Equation (VI.1). Relationship neither within nor between base clusterings (i.e., bc_{j_1} and bc_{j_2}) is explicated. Besides, most existing work [1], [8], [3] only uses the similarity measure between objects when clustering them. It thus does not involve the context (i.e., neighborhood) of the objects.

To solve the first two issues above, we regard the entry $BC^*(x,y)$ of the overall similarity matrix to be *IaOSO*:

$$S_O^{IaC}(u_x, u_y) = BC^*(x,y) = \delta^{IaO}(u_x, u_y), \tag{VI.8}$$

where δ^{IaO} is defined in (V.5). Here, S_O^{IaC} captures the intra-coupled interactions within two objects as well as both the intra-coupled and inter-coupled interactions among base clusterings. Alternatively, we can also assign $BC_j(x,y)$ of base clustering bc_j to be δ_j^C (V.4), in the same way as Equation (VI.3); then, the overall similarity matrix BC^* is obtained by averaging the associated similarity matrix BC_j over all the base clusterings according to (VI.7). Afterwards, *METIS* is applied to the overall similarity matrix BC^* to produce the final clustering fc^*. We denote this newly proposed intra-coupled object-based coupling method as *OC-Ia*.

Further considering the above third issue, both the intra-couplings and inter-couplings of clusterings and of objects are incorporated as follows:

$$S_O^C(u_x, u_y) = BC^*(x,y) = \delta^{CO}(u_x, u_y | U), \tag{VI.9}$$

where δ^{CO} is defined in (V.9). Since we would like to maximize the sum of $\delta^{CO}(u_x, u_y|U)$ (V.9) for data object pairs u_x, u_y belonging to a single cluster, and at the same time, minimize the sum of $\delta^{CO}(u_x, u_y|U)$ for u_x and u_y in different clusters. Accordingly, the desired final clustering $fc^* = \{c_*^1, \cdots, c_*^{t^*}\}$ with t^* clusters can be obtained by maximizing the following criterion function:

$$fc^* = \underset{c^1,\cdots,c^{t^*}}{\arg} \max \sum_{t=1}^{t^*} m_t \cdot \sum_{u_x, u_y \in c^t} \frac{S_O^C(u_x, u_y) \cdot m}{m_t^{1+2f(\theta)}}, \tag{VI.10}$$

where c^t denotes the tth cluster of size m_t, m is the total number of objects, and $f(\theta) = (1-\theta)/(1+\theta)$. The rationale of the above function is twofold: on one hand, one of our goals is to maximize $\delta^{CO}(u_x, u_y|U)$ for all pairs of objects in the same cluster $u_x, u_y \in c^t$; on the other hand, we divide the total *CCOSO* (i.e., $S_O^C = \delta^{CO}$) involving pairs of objects in cluster c^t by the expected sum of δ^{CO} in c^t, which is $m_t^{1+2f(\theta)}/m$ [6]; and then weigh this quantity by m_t, i.e., the number of objects in c^t. Dividing by the expected sum of δ^{CO} prevents the case of a clustering in which all objects are assigned to a single cluster and objects with very small coupled similarity value between them from being put in the same cluster [6]. Subsequently, we adapt the standard agglomerative hierarchical clustering algorithm to obtain the final clustering

fc^* by solving Equation (VI.10) [6]. We abbreviate this newly proposed hierarchical object-based coupling to be *OC-H*.

C. Cluster-based Coupling

The cluster-based consensus function characterizes the interactions between every two clusters. One of the basic approaches based on the relationship between clusters is *MCLA* proposed by Strehl and Ghosh [2]. The idea in *MCLA* is to yield object-wise confidence estimates of cluster membership, to group and then to collapse related clusters represented as hyperedges. The similarity measure of clusters in *MCLA* is Jaccard matching coefficient [15], formally:

$$S_{Cr}(c_{j_1}^{t_1}, c_{j_2}^{t_2}) = \frac{|c_{j_1}^{t_1} \cap c_{j_2}^{t_2}|}{|c_{j_1}^{t_1} \cup c_{j_2}^{t_2}|}, \quad \text{(VI.11)}$$

where $c_{j_1}^{t_1}$ and $c_{j_2}^{t_2}$ are the t_1th cluster of base clustering bc_{j_1} and the t_2th cluster of base clustering bc_{j_2}, respectively.

The above similarity measure S_{Cr} considers neither coupling between base clusterings nor interaction between objects. Therefore, it is in lack of the capability to reflect the essential link and relationship among data. In order to remedy this problem, we define the coupled similarity between clusters $c_{j_1}^{t_1}$ and $c_{j_2}^{t_2}$ in terms of both the coupled relationships between clusterings and between objects. The average sum of every two-object pairs in $c_{j_1}^{t_1}$ and $c_{j_2}^{t_2}$ respectively is selected here to specify the coupled similarity between clusters:

$$S_{Cr}^{C}(c_{j_1}^{t_1}, c_{j_2}^{t_2}) = \frac{1}{m_{t_1} m_{t_2}} \sum_{u_x \in c_{j_1}^{t_1}, u_y \in c_{j_2}^{t_2}} S_O(u_x, u_y), \quad \text{(VI.12)}$$

where m_{t_1} and m_{t_2} are the sizes of clusters $c_{j_1}^{t_1}$ and $c_{j_2}^{t_2}$, respectively; $S_O(u_x, u_y)$ is the coupled similarity for objects, it can be either δ^{IaO} (V.5) or δ^{CO} (V.9). If $S_O = \delta^{IaO}$, the cluster-based coupling includes the intra and inter-coupled interaction between base clusterings as well as the intra-coupled interaction between objects; if $S_O = \delta^{CO}$, it reveals both the intra and inter-coupled interactions between base clusterings and between objects. Afterwards, *METIS* is used based on the cluster-cluster similarity matrix to conduct meta-clustering as in [2]. We denote the cluster-based coupling as *CrC* (including *CrC-Ia* with δ^{IaO} and *CrC-C* with δ^{CO}).

D. Miscellaneous Issues

How to Generate Base Clusterings: There are several existing methods to provide diverse base clusterings: using different clustering algorithms, employing random or different parameters of some algorithms, and adopting random subsampling or random projection of the data. Since our focus is mainly on the consensus function, we use *k-means* on random subsampling [8] of the data as the base clustering algorithm in our experiments. The number t^j of base clustering bc_j is pre-defined for each data set and remains the same for all clustering runs.

How to Post-process Clustering: In the proposed *CCE* framework, we mainly focus on the consensus function based on pairwise interactions between base clusterings, between objects and between clusters. Those interactions are described by the corresponding similarity matrices. Thus, a common and recommended way to combine the base clusterings is to recluster the objects using any reasonable similarity-based clustering algorithm. In our experiments, we choose *k-means*, *agglomerative algorithm* [6] and *METIS* [2] due to their popularity in clustering ensemble.

VII. Algorithm and Analysis

In previous sections, we have discussed the coupled framework of clustering ensembles *CCE* from the perspectives of coupling of clusterings, coupling of objects, and coupled consensus functions. They are all based on the intra and inter-coupled interactions between clusterings and between objects. Therefore, in this section, we design two algorithms *CCSC*[1] (Algorithm 1) and *CCOSO* (Algorithm 2) to compute the coupled similarity for each pair of cluster labels and the coupled similarity for objects u_x and u_y, respectively.

Algorithm 1: Coupled Similarity for Clusters *CCSC*

Data: Object set $U = \{u_1, \cdots, u_m\}$ and $u_x, u_y \in U$, base clustering set $C = \{bc_1, \cdots, bc_L\}$, and weight $\lambda = (\lambda_k)_{1\times L}$.
Result: Similarity matrix *CCSC* between cluster labels.

begin
  maximal cluster label $r(j) \longleftarrow \max(V_j)$
  for *every cluster label pair* $(v_j^x, v_j^y \in [1, r(j)])$ **do**
    $U_1 \longleftarrow \{i | v_j^i == v_j^x\}$, $U_2 \longleftarrow \{i | v_j^i == v_j^y\}$
    // Compute intra-coupled similarity between cluster labels v_j^x and v_j^y.
    $\delta_j^{IaC}(v_j^x, v_j^y) = (|U_1||U_2|)/(|U_1| + |U_2| + |U_1||U_2|)$
    $\delta_j^{C}(v_j^x, v_j^y | \{V_k\}_{k=1}^{L}) \longleftarrow$
    $\delta_j^{IaC}(v_j^x, v_j^y) \cdot IeCSC(v_j^x, v_j^y)$
  $CCSC(v_j^x, v_j^y) \longleftarrow \delta_j^{C}(v_j^x, v_j^y | \{V_k\}_{k=1}^{L})$
end

Function $IeCSC(v_j^x, v_j^y, U_1, U_2)$
begin
  for *each base clustering* $(bc_k \in C) \wedge (bc_k \neq bc_j)$ **do**
    $\varphi \longleftarrow \{v_k^x | x \in U_1\} \cap \{v_k^y | y \in U_2\}$
    for *every intersection* $v_k^z \in \varphi$ **do**
      $U_0 \longleftarrow \{i | v_k^i == v_k^z\}$
      $ICP_x \longleftarrow |U_0 \cap U_1|/|U_1|$
      $ICP_y \longleftarrow |U_0 \cap U_2|/|U_2|$
      $Min_{(x,y)} \longleftarrow min(ICP_x, ICP_y)$
    $\delta_{j|k}(v_j^x, v_j^y | V_k) = sum[Min_{(x,y)}]$
  // Compute inter-coupled similarity between two cluster labels v_j^x and v_j^y.
  $\delta_j^{IeC}(v_j^x, v_j^y | \{V_k\}_{k\neq j}) = sum[\lambda_k \cdot \delta_{j|k}(v_j^x, v_j^y | V_k)]$
  return $IeCSC(v_j^x, v_j^y) = \delta_j^{IeC}(v_j^x, v_j^y | \{V_k\}_{k\neq j})$

As shown in these two algorithms, the computational complexity for *CCSC* is $O(LT^3)$, and the computational complexity for *CCOSO* is $O(L^2T^3 + 2m)$, where L is the number of

[1]All the cluster labels of each base clustering need to be encoded as numbers, starting at one and increasing to the maximum which is the respective number of clusters in this base clustering.

base clusterings, T is the maximal number of clusters in all the base clusterings, and m is the total number of objects.

Algorithm 2: Coupled Similarity for Objects *CCOSO*

Data: Object set $U = \{u_1, \cdots, u_m\}$ and $u_x, u_y \in U$, base clustering set $C = \{bc_1, \cdots, bc_L\}$, and threshold $\theta \in [0, 1]$.
Result: Similarity $CCOSO(u_x, u_y)$ between objects u_x, u_y.

```
begin
    for each base clustering bc_j ∈ C do
        δ_j^C(v_j^x, v_j^y | {V_k}_{k=1}^L) ⟵ CCSC(v_j^x, v_j^y)
    // Compute intra-coupled similarity
       between two objects u_x and u_y.
    δ^{IaO}(u_x, u_y) = 1/L · sum[δ_j^C(v_j^x, v_j^y | {V_k}_{k=1}^L)]
    neighbor sets N_{u_x} = N_{u_y} = ∅
    for objects (u_{z_1}, u_{z_2} ∈ U) ∧ (u_{z_1} ≠ u_x) ∧ (u_{z_2} ≠ u_y) do
        if δ^{IaO}(u_x, u_{z_1}) ≥ θ then
            N_{u_x} = {u_{z_1}} ∪ N_{u_x}
        if δ^{IaO}(u_y, u_{z_2}) ≥ θ then
            N_{u_y} = {u_{z_2}} ∪ N_{u_y}
    // Compute inter-coupled similarity
       between two objects u_x and u_y.
    δ^{CO}(u_x, u_y | U) = 1/m · |N_{u_x} ∩ N_{u_y}|
    CCOSO(u_x, u_y) ⟵ δ^{CO}(u_x, u_y | U)
end
```

VIII. Empirical Study

This section presents the performance evaluation of the coupled framework *CCE* in terms of the clustering-based (*CgC*), object-based (*OC-Ia* and *OC-H*), and cluster-based (*CrC-Ia* and *CrC-C*) couplings. The experiments are conducted on 11 synthetic and real data sets to validate accuracy and stability of various consensus functions.

A. Data Sets

The experimental evaluation is conducted on eight data sets, including two synthetic data sets (i.e., Sy1 and Sy2, which are 2-Gaussian and 4-GaussianN, respectively) and nine real-life data sets from UCI. Table III summarizes the details of these data sets, where m is the number of objects, n is the number of dimensions, and t^p is the number of pre-known classes. Those true classes are only used to evaluate the quality of the clustering results, not in the process of aggregating base clusterings. The number of true classes is only used to set the number of clusters both in building the base clusterings and in the post-processing stage. Since we do not involve the information of attributes after building base clusterings, we order the data sets according to the number of objects ranging from 150 to 1484. Note that the second synthetic data set Sy2 [16] is initially created in two dimensions and later added with four more dimensions of uniform random noise.

B. Selection of Parameters and Algorithms

As previously presented, our experiments are designed from the following three perspectives:

TABLE III
Description of Data Sets

Data Set	m	n	t^p	Source
Sy1	200	2	2	modified from [2]
Sy2	400	6	4	modified from [16]
Iris	150	4	3	UCI repository
Wine	178	13	3	UCI repository
Seg	210	19	7	UCI repository
Glass	214	9	6	UCI repository
Ecoli	336	7	8	UCI repository
Ionos	351	34	2	UCI repository
Blood	748	5	2	UCI repository
Vowel	990	10	11	UCI repository
Yeast	1484	8	10	UCI repository

1) Clustering-based: Besides the partition difference (*PD*) proposed in [9], *QMI* is also an effective clustering-based criterion [4], which has proved to be equivalent with *Category Utility Function* in [9]. We will compare the clustering-based coupling (*CgC*) with its baseline method *PD* [9], *EM* and *QMI* [4].
2) Object-based: In this group, we will compare the intra-coupled object-based coupling *OC-Ia* with its baseline method *CSPA* [2], and compare the hierarchical object-based coupling *OC-H* with *CSPA* [2] and with the categorical clustering algorithms: *ROCK* [6] (the baseline method of *OC-H*) and *LIMBO* [12].
3) Cluster-based: Based on *MCLA* [2], *HBGF* is another promising cluster-based criterion [8]. It also collectively considers the similarity between objects and clusters but lacks the discovery of coupling. Iam-On et al. [5] proposed a link-based approach (*LB*), which is an improvement on *HBGF*. Below, cluster-based coupling *CrC* (including *CrC-Ia* and *CrC-C*) is compared with their baseline method *MCLA* [2], *HBGF* [8], and *LB* [5] (including *LB-P* and *LB-S*).

As indicated in Section VI-D, *k-means* on random subsampling [8] of the data is used to produce a diversity of base clusterings; *k-means* and *agglomerative algorithm* are used to post-process the coupled consensus functions *CgC* and *OC-H*, respectively, and *METIS* is adopted to post-process the consensus functions *OC-Ia*, *CrC-Ia* and *CrC-C*. The following parameters of the clustering ensemble are especially important:

- θ: The neighbor threshold in (V.6) is defined to be the average *IaOCO* and Jaccard coefficient [6] values of pairwise objects for *OC-H* and *ROCK*, respectively.
- L: The ensemble size (i.e., the number of base clusterings) is taken to be $L = 10$.
- t^j, t^*: The number of clusters in the base clustering bc_j and final clustering fc^* are both regarded as the number of pre-known classes t^p, i.e., $t^j = t^* = t^p$.
- λ_k: The weight λ_k for base clustering bc_k in Definition 5.3 on *IeCSC* is simplified as $\lambda_k = 1/L = 1/10$.
- NR: The number of runs for each clustering ensemble is fixed to be $NR = 50$ to obtain corresponding average results for the evaluation measures below.

Other parameters of the compared methods remain the same as the original approaches.

TABLE IV
EVALUATION MEASURES ON BASE CLUSTERINGS

Data Set	AC			NMI			CSI
	Max	Avg	Min	Max	Avg	Min	Avg
Sy1	0.955	0.950	0.945	0.745	0.720	0.693	0.714
Sy2	0.503	0.460	0.385	0.406	0.406	0.406	0.698
Iris	0.927	0.827	0.513	0.750	0.656	0.427	0.791
Wine	0.708	0.689	0.556	0.441	0.424	0.388	0.659
Seg	0.586	0.529	0.433	0.548	0.496	0.410	0.820
Glass	0.517	0.479	0.449	0.338	0.307	0.276	0.602
Ecoli	0.687	0.512	0.470	0.539	0.437	0.398	0.530
Ionos	0.712	0.704	0.650	0.131	0.107	0.014	0.670
Blood	0.739	0.709	0.707	0.017	0.016	0.013	0.780
Vowel	0.373	0.354	0.339	0.435	0.415	0.388	0.802
Yeast	0.384	0.332	0.319	0.250	0.220	0.218	0.817

Since each clustering ensemble method divides data objects into a partition of t^p (i.e., the number of true classes) clusters, we then evaluate the clustering quality against the corresponding true partitions by using these external criteria: accuracy (AC) [13], normalized mutual information (NMI) [13], and combined stability index (CSI) [16]. AC and NMI describe the degree of approximation between obtained clusters and the true data classes. CSI reveals the stability between them across $NR = 50$ runs, it reflects the deviation of the results across different runs. In fact, the larger AC or NMI or CSI is, the better the clustering ensemble algorithm is. Note that the correspondence problem between the derived clusters and the known classes need to be solved before evaluation. The optimal correspondence can be obtained using the Hungarian method [4] with $O((t^p)^3)$ complexity for t^p clusters.

C. Experimental Results

Based on the evaluation measures (i.e., AC, NMI and CSI), Table IV displays the performance of the base clustering algorithm (i.e., *k-means*) over synthetic and real data sets. Note that Max, Avg, and Min represent the maximal, average, and minimum corresponding evaluation scores among input base clusterings, respectively.

In the following, the experimental results are presented and analyzed in three groups: clustering-based comparison which focuses on the evaluation of coupling between base clusterings, object-based comparison which studies the utility of intra-coupling and inter-coupling between objects, and cluster-based comparison which identifies the jointed effect of couplings both between base clusterings and between objects. We individually analyze the clustering performance by considering the couplings step by step within each group of experiments, and the comparison across these three groups is out of scope of this paper. Note that all the values reported on AC and NMI are the averages across multiple clustering ensembles (i.e., exactly 50 runs), CSI value reveals the total deviation apart from the average of 50 runs in each experiment, and the improvement rate below refers to the absolute difference value between two evaluation scores.

Clustering-based Comparison: Fig. 5 shows the performance comparison of different clustering-based ensemble methods over two synthetic and six real-life data sets in terms

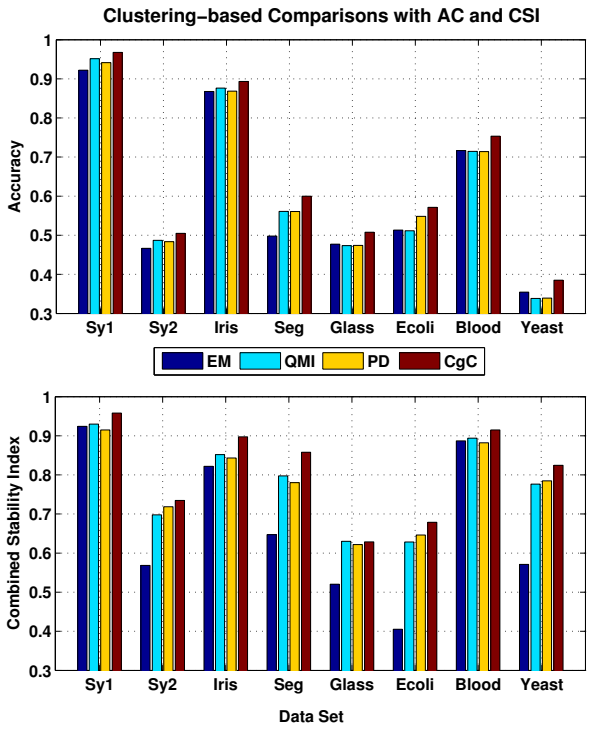

Fig. 5. Clustering-based comparisons.

of AC and CSI. Due to the space limitation, the results on only eight data sets are reported; and the performance on NMI, which is similar to that on AC, is not provided here. It is clear that our proposed *CgC* usually generates data partitions of higher quality than its baseline model *PD* and other compared approaches, i.e., *EM* and *QMI*. Specifically, in terms of accuracy, the AC improvement rate ranges from 1.59% (*QMI* on Sy1) to 10.20% (*EM* on Seg), and there has been significant CSI improvement (from 0.69% to 27.35%) except one case: Glass. Overall, the average improvement rate for AC across all the methods over every data set is 3.71%, and the average improvement rate of *CgC* on CSI is 7.26%. Also, in several data sets such as Sy1, Sy2, Seg, Blood and Yeast, their AC measures exceed the maximum of AC in the corresponding base clusterings, i.e., Max(AC) in Table IV. Besides, all the AC and CSI values of *CgC* are higher than the corresponding average values of base clustering. Another observation is that none of the other three compared consensus functions is the absolute winner among them, while *QMI* is the best in most cases, followed by *PD* and with *EM* to be the worst one. But our proposed *CgC* outperforms all the compared algorithms on almost every data set. The improvement level is associated with the accuracy of base clusterings: the higher accuracy of base clusterings corresponds to relative smaller level of improvement. Statistical analysis, namely t-test, has been done on the AC of our *CgC*, with 95% confidence level. The null hypothesis that *CgC* is better than base clusterings and the best result of other methods in terms of AC is accepted.

Therefore, we obtain the empirical conclusion that clustering accuracy and stability can be further improved with *CgC* by involving the couplings of clusterings. The improvement

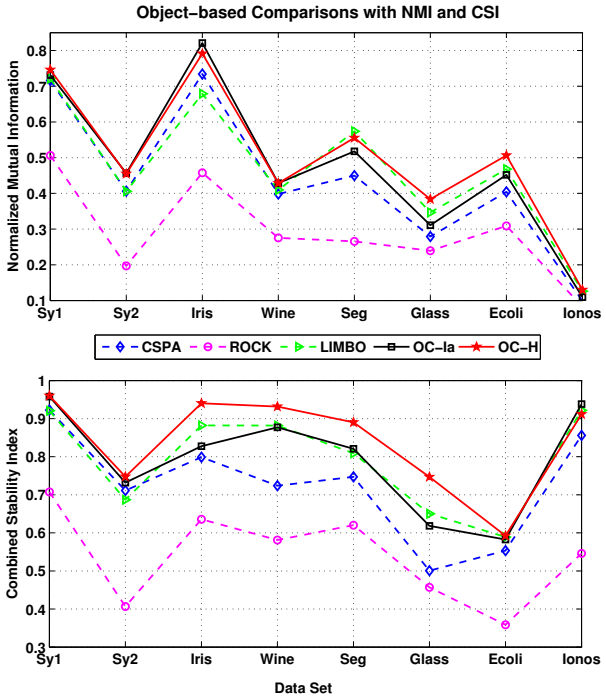

Fig. 6. Object-based comparisons.

rate is dependent on the accuracy of base clusterings.

Object-based Comparison: The evaluation (i.e., NMI and CSI) of distinct object-based ensemble methods are exhibited in Fig. 6. Eight data sets with smaller size are chosen because of the high computational complexity in this group of experiments. Due to the space limitation, the performance on AC, which is similar to that on NMI, is not reported here. We observe that, with the exception of a few items, our proposed *OC-Ia* mostly outperforms the ensemble method *CSPA* and categorical clustering algorithm *ROCK* in terms of both NMI and CSI. Our proposed *OC-H* has the largest NMI and CSI values over most of the data sets. Here, it can be clearly seen that our proposed *OC-Ia* and *OC-H* both achieve better clustering quality when compared to their respective baseline methods *CSPA* and *ROCK*. The average NMI and CSI improvement rates for the former pair are 4.25% and 6.76%, respectively, and those values for the latter pair are 20.80% and 30.10%. When compared with Table IV, all the NMI and CSI values of *OC-Ia* and *OC-H* are greater than the corresponding average values of base clustering, and several NMI values are even larger than that of the maximum in base clustering, e.g., Sy2 and Iris. It is also noteworthy that the evaluation scores of categorical clustering algorithm *LIMBO* are comparable with our proposed *OC-Ia*, but worse than *OC-H*. The reason is that *LIMBO* also considers the coupling between attributes from the perspective of information theory, but without the concern of the coupling between objects. However, also as a categorical clustering algorithm, *ROCK* leads to a poor performance in the clustering ensemble, since it only focuses on the interaction between objects but overlooks the relationship between base clusterings. Statistical test supports the results on NMI.

Thus, the involvement of intra-coupling between objects (e.g., *OC-Ia*) and inter-coupling between objects (e.g., *OC-H*) can both enhance the clustering quality, while the latter one performs a bit better.

Cluster-based Comparison: Table V reports the experiment results with the cluster-based ensemble methods by using evaluation measures: AC, NMI and CSI. The two highest measure scores of each experimental setting are highlighted in boldface. The last column is the average value for associated measures across all the data sets. As this table indicates, our proposed *CrC-Ia* and *CrC-C* mostly get the first two positions on every individual data set, and their average evaluation scores are the corresponding largest two among all the average values. For AC, the average improvement rate of *CrC-Ia* and *CrC-C* against other methods ranges from 1.84% to 6.79%; for NMI, the minimal and maximal average improvement rates are 2.19% and 6.56%, respectively; for CSI, this rate falls between 2.02% and 12.44%. Resembling the above comparisons, all the evaluation scores of *CrC-Ia* and *CrC-C* are at least not smaller than the corresponding average values of base clustering, with several AC and NMI values even greater than the relevant maximal scores in base clustering, e.g., Sy2 and Wine. Another significant observation is that the average AC and NMI improvement rates of *CrC-C* on *CrC-Ia* are only 1.86% and 1.42% respectively, which are smaller than those of *CrC-Ia* and *CrC-C* on other compared methods. We know that *CrC-C* built on *CrC-Ia* also involves the common neighborhood of objects. When most of the base clusterings have a relative consistent grouping of the objects, the chance to encounter the situation that half of the base clusterings put two objects in the same cluster while the rest half separate them in different groups is rare. Therefore, the improvement made by *CrC-C* upon *CrC-Ia* is minor or even negative in this scenario, such as Seg and Yeast whose CSI values across 10 base clusterings are as high as 0.820 and 0.817 in Table IV, respectively. However, for a majority of cases, different base clusterings result in various results. Thus, *CrC-C* is expected to have a better performance in particular when differentiating those questionable objects, compared to *CrC-Ia*. All the results on AC and NMI are supported by a statistical significant test with a confidence level at 95%.

Consequently, the clustering quality benefits from both the couplings between clusterings and between objects. However, the inter-coupling of objects is dependent on the consistency of base clustering results, which affects the improvement degree.

We draw the following three conclusions to address the research questions proposed in Section I: 1) Base clusterings are indeed coupled with each other, and the consideration of such couplings can result in better clustering quality; 2) The inclusion of coupling between objects further improves the clustering accuracy and stability; 3) The improvement level brought by the coupling of base clusterings is associated with the accuracy of base clusterings, while the improvement degree caused by the inter-coupling of objects is dependent on the consistency of base clustering results.

TABLE V
CLUSTER-BASED COMPARISONS ON AC, NMI AND CSI

Data Set		Sy1	Sy2	Iris	Wine	Seg	Glass	Ecoli	Ionos	Blood	Vowel	Yeast	Avg
AC	*MCLA*	0.945	0.501	0.875	0.702	0.560	0.472	0.528	0.711	0.680	0.365	0.341	0.607
	HBGF	0.949	0.503	0.877	0.690	0.532	0.445	0.468	0.684	0.528	0.379	0.301	0.578
	LB-P	0.952	0.504	0.878	0.703	**0.582**	0.459	0.530	0.711	**0.719**	0.330	0.328	0.609
	LB-S	0.951	0.486	0.844	0.690	0.560	**0.483**	**0.539**	0.711	0.713	0.364	0.332	0.607
	CrC-Ia	**0.954**	**0.513**	**0.893**	**0.731**	**0.579**	0.482	**0.539**	**0.721**	0.713	**0.394**	**0.379**	**0.627**
	CrC-C	**0.969**	**0.518**	**0.902**	**0.764**	**0.579**	**0.511**	**0.587**	**0.742**	**0.723**	**0.430**	**0.378**	**0.646**
NMI	*MCLA*	0.725	0.406	0.744	0.429	0.526	0.318	0.510	0.129	0.015	0.411	0.223	0.403
	HBGF	0.710	0.389	0.706	0.355	0.486	0.316	0.444	0.109	0.007	0.414	0.206	0.377
	LB-P	0.723	0.406	0.745	0.429	**0.548**	0.318	**0.511**	0.130	0.016	0.420	0.221	0.406
	LB-S	0.724	0.363	0.687	0.412	0.531	**0.335**	0.502	0.130	0.015	0.394	0.210	0.391
	CrC-Ia	**0.734**	**0.436**	**0.752**	**0.556**	**0.543**	0.323	**0.511**	**0.164**	**0.018**	**0.445**	**0.226**	**0.428**
	CrC-C	**0.764**	**0.456**	**0.753**	**0.580**	0.540	**0.337**	**0.539**	**0.171**	**0.019**	**0.477**	**0.228**	**0.442**
CSI	*MCLA*	0.950	0.710	0.876	0.828	0.775	0.554	0.640	0.937	**0.897**	0.783	0.774	0.793
	HBGF	0.953	0.703	0.761	0.712	0.716	0.594	0.528	0.839	0.642	0.736	0.742	0.721
	LB-P	0.954	0.713	0.860	0.829	0.840	0.601	**0.673**	0.943	0.893	0.774	0.786	0.806
	LB-S	0.943	0.662	0.787	0.846	0.767	0.601	0.594	0.926	0.892	0.757	0.727	0.773
	CrC-Ia	**0.967**	**0.736**	**0.892**	**0.868**	**0.878**	**0.621**	0.649	**0.955**	**0.897**	**0.808**	**0.817**	**0.826**
	CrC-C	**0.963**	**0.752**	**0.910**	**0.880**	**0.880**	**0.639**	**0.679**	**0.957**	**0.940**	**0.872**	**0.822**	**0.845**

IX. CONCLUSION AND FUTURE WORK

Clustering ensemble has been introduced as a more accurate alternative to individual (base) clustering algorithms. However, the existing approaches mostly assume the independency between base clusterings, and overlook the coupling between objects in terms of neighborhood. This paper proposes a novel framework for coupled clustering ensembles, i.e. *CCE*, to incorporate interactions both between base clusterings and objects. *CCE* caters for cluster label frequency distribution within one base clustering (intra-coupling of clusterings), cluster label co-occurrence dependency between distinct base clusterings (inter-coupling of clusterings), base clustering aggregation between two objects (intra-coupling of objects), and neighborhood relationship among other objects (inter-coupling of objects), which has been shown to improve learning accuracy and stability. Substantial experiments have verified that the consensus functions incorporated with these couplings significantly outperform nine state-of-art techniques in terms of clustering-base, object-based and cluster-based ensembles as well as the algorithm to produce base clusterings (*k-means*).

This work verifies that the couplings between clusterings and between objects are essential to the clustering ensemble problem. The coupling of clusterings can enhance the clustering quality in most cases, and the improvement level depends on the accuracy of base clusterings. The inter-coupling of objects is associated with the consistency of base clustering results, which leads to fluctuated improvement on the clustering quality. Thus, what is the relationship between the coupling of base clusterings and their individual performance? What is the relationship between the coupling of objects and the consistency? How about introducing a weight to control the coupling of objects during the process of clustering ensemble? How to fix the weights λ_k of base clustering bc_k in *IeCSC* rather than simply treating them to be equal? We are currently working on these potential issues, and will involve the coupling of clusters in our future work.

ACKNOWLEDGMENT

This work is sponsored in part by Australian Research Council Discovery Grant (DP1096218) and Australian Research Council Linkage Grant (LP100200774).

REFERENCES

[1] I. Christou, "Coordination of cluster ensembles via exact methods," *IEEE TPAMI*, vol. 33, no. 2, pp. 279–293, 2011.
[2] A. Strehl and J. Ghosh, "Cluster ensembles–a knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, 2002.
[3] A. Gionis, H. Mannila, and P. Tsaparas, "Clustering aggregation," in *ICDE 2005*, 2005, pp. 341–352.
[4] A. Topchy, A. Jain, and W. Punch, "Clustering ensembles: models of consensus and weak partitions," *IEEE TPAMI*, vol. 27, no. 12, pp. 1866–1881, 2005.
[5] N. Iam-On, T. Boongoen, S. Garrett, and C. Price, "A link-based approach to the cluster ensemble problem," *IEEE TPAMI*, vol. 33, no. 12, pp. 2396–2409, 2011.
[6] S. Guha, R. Rastogi, and K. Shim, "ROCK: A robust clustering algorithm for categorical attributes," *Information Systems*, vol. 25, no. 5, pp. 345–366, 2000.
[7] L. Cao, Y. Ou, and P. S. Yu, "Coupled behavior analysis with applications," *IEEE TKDE*, vol. 24, no. 8, pp. 1378–1392, 2011.
[8] X. Z. Fern and C. E. Brodley, "Solving cluster ensemble problems by bipartite graph partitioning," in *ICML 2004*, 2004, pp. 36–43.
[9] T. Li, M. Ogihara, and S. Ma, "On combining multiple clusterings: an overview and a new perspective," *Applied Intelligence*, vol. 33, no. 2, pp. 207–219, 2010.
[10] K. Punera and J. Ghosh, "Soft cluster ensembles," *Advances in fuzzy clustering and its applications*, pp. 69–91, 2007.
[11] C. Domeniconi and M. Al-Razgan, "Weighted cluster ensembles: methods and analysis," *TKDD*, vol. 2, no. 4, p. 17, 2009.
[12] P. Andritsos, P. Tsaparas, R. Miller, and K. Sevcik, "LIMBO: Scalable clustering of categorical data," in *EDBT 2004*, 2004, pp. 123–146.
[13] C. Wang, L. Cao, M. Wang, J. Li, W. Wei, and Y. Ou, "Coupled nominal similarity in unsupervised learning," in *CIKM 2011*, 2011, pp. 973–978.
[14] S. Cost and S. Salzberg, "A weighted nearest neighbor algorithm for learning with symbolic features," *Machine Learning*, vol. 10, no. 1, pp. 57–78, 1993.
[15] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*. SIAM, Philadelphia, ASA, Alexandria, VA: ASA-SIAM Series on Statistics and Applied Probability, 2007.
[16] L. Kuncheva and D. Vetrov, "Evaluation of stability of k-means cluster ensembles with respect to random initialization," *IEEE TPAMI*, vol. 28, no. 11, pp. 1798–1808, 2006.