Coupled Clustering Ensemble by Exploring Data Interdependence

CAN WANG, School of Information and Communication Technology, Griffith University, Australia.
CHI-HUNG CHI, Data61, Commonwealth Scientific and Industrial Research Organisation, Australia.
ZHONG SHE, School of Information and Communication Technology, Griffith University, Australia.
LONGBING CAO, Advanced Analytics Institute, University of Technology, Sydney, Australia.
BELA STANTIC, School of Information and Communication Technology, Griffith University, Australia.

Clustering ensembles combine multiple partitions of data into a single clustering solution. It is an effective technique for improving the quality of clustering results. Current clustering ensemble algorithms are usually built on the pairwise agreements between clusterings that focus on the similarity via consensus functions, between data objects that induce similarity measures from partitions and re-cluster objects, and between clusters that collapse groups of clusters into meta-clusters. In most of those models, there is a strong assumption on IIDness (i.e. independent and identical distribution), which states that base clusterings perform independently of one another and all objects are also independent. In the real-world, however, objects are generally likely related to each other through features that are either explicit or even implicit. There is also latent but definite relationship among intermediate base clusterings because they are derived from the same set of data. All these demand a further investigation of clustering ensembles that explores the interdependence characteristics of data. To solve this problem, a new coupled clustering ensemble (i.e. *CCE*) framework that works on the interdependence nature of objects and intermediate base clusterings is proposed in this paper. The main idea is to model the coupling relationship between objects by aggregating the similarity of base clusterings, and the interactive relationship among objects by addressing their neighborhood domains. Once these interdependence relationships are discovered, they will act as critical supplements to clustering ensembles. We verified our proposed framework by using three types of consensus function: clustering-based, object-based, and cluster-based. Substantial experiments on multiple synthetic and real-life benchmark data sets indicate that *CCE* can effectively capture the implicit interdependence relationships among base clusterings and among objects with higher clustering accuracy, stability, and robustness compared to 14 state-of-the-art techniques, supported by statistical analysis. In addition, we show that the final clustering quality is dependent on the data characteristics (e.g. quality and consistency) of base clusterings in terms of sensitivity analysis. Finally, the applications in document clustering, as well as on the data sets with much larger size and dimensionality, further demonstrate the effectiveness, efficiency, and scalability of our proposed models.

CCS Concepts: •**Computing methodologies** → **Ensemble methods;** *Learning latent representations;* •**Information systems** → **Clustering;**•**Applied computing** → Document analysis;

Additional Key Words and Phrases: clustering ensemble, behavior interior dimensions, interdependence, base clustering, object, coupling.

1. INTRODUCTION

Clustering analysis is a fundamental tool for capturing the structure of data. Lots of clustering algorithms [Kriegel et al. 2009; Havens et al. 2012; Shao et al. 2016] have been proposed, but the No Free Lunch theorem [Wolpert and Macready 1996] suggests that there is no single, supreme algorithm that fits all cluster shapes and structures perfectly. Consequently, as a recent offshoot of classifier ensemble research [García-Osorio et al. 2010; Sun et al. 2015], the clustering ensemble [Vega-Pons and Ruiz-Shulcloper 2011; Franek and Jiang 2014] has exhibited great potential for

This work was supported in part by Griffith University's 2018 New Researcher Grant.

Author's addresses: C. Wang, Z. She and B. Stantic, School of Information and Communication Technology, Griffith University, Gold Coast, QLD 4215, Australia; C-H. Chi, Data61, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Sandy Bay, TAS 7005, Australia; L. Cao, Advanced Analytics Institute, University of Technology, Sydney, NSW 2007, Australia.
Email: can.wang@griffith.edu.au (Corresponding Author: Can Wang).

DOI: http://dx.doi.org/10.1145/0000000.0000000

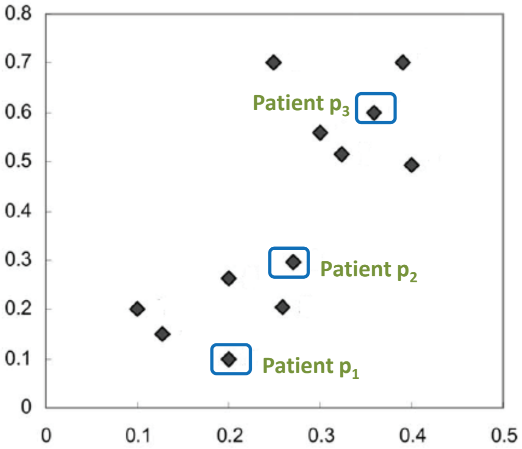


Fig. 1. 2D nonclassical multidimensional scaling of indicators for 12 patients.

enhancing clustering accuracy, robustness and parallelism [Strehl and Ghosh 2002; Gan and Ng 2015] by combining results from various clustering methods.

In general, the clustering ensemble process can be divided into three stages: building base clusterings, aggregating base clusterings, and post-processing clustering. The ultimate objective is to produce an overall high-quality clustering that agrees as much as possible with each of the input clusterings. The essence of the clustering ensemble is to aggregate the advantages of each base clustering to give a more complete, global understanding of the underlying data, assuming that each base clustering is to capture the best local picture of the same data set. While the clustering ensemble often captures the common structure of base clusterings and achieves better clustering quality than that of individual ones, there are still open challenges that have not been well explored in the consensus design, in particular the one related to the IIDness (i.e. independent and identical distribution) of data. We illustrate the problem and discuss the challenges of clustering ensemble.

Fig. 1 shows the data of a 2D non-classical multidimensional scaling of testing indicators for 12 patients. Each patient is originally described by 13 attributes, such as age, sex, chest pain type, serum cholesterol, etc. The objective to analyze these indicators is to justify who might be attacked by heart disease, and then recommend proactive treatments for those patients. Accordingly, we apply k-means to do grouping on this data set which has 12 objects, each with 13 attributes. The predefined number of clusters is set to two because we are interested in the presence and absence of heart disease. In each run, we obtain different clustering results since k-means is rather sensitive to the random initialization. As can be shown in Table I, for patients p_1, p_2, and p_3 for instance, there are four distinct groupings resulted from four base clusterings (bc_1, bc_2, bc_3, and bc_4). The target of the clustering ensemble is to obtain a final clustering based on these four base clusterings.

Table I shows the four possible cluster labels for p_1, p_2, and p_3. It shows that the first two base clusterings, namely bc_1 and bc_2, assign p_1 and p_2 in the same cluster while the last two base clusterings, bc_3 and bc_4, assign p_1 and p_2 to different clusters. This creates problem in the last stage of post-processing clustering because using traditional clustering ensemble such as CSPA [Strehl and Ghosh 2002], the similarity of each pair of the three objects, p_1, p_2, and p_3, are all the same (i.e. $Sim(p_1, p_2) = Sim(p_1, p_3) = Sim(p_2, p_3) = 0.5$). The root of this problem is the assumption of "IIDness" assumption [Cao 2014]. The consensus building assumes that all four base clusterings are independent, and that each base clustering also treats all the patients independently. To solve this dilemma, one common approach is to assign them randomly either in the same cluster or in different clusters, and this is clearly subject to questions because of the randomness of allocation.

Table I. Base Clustering Results for Three Patients

	bc_1	bc_2	bc_3	bc_4
Patient p_1	2	A	X	α
Patient p_2	2	A	Y	β
Patient p_3	1	A	Y	α

If we study Table I carefully, we find that all the three patients p_1, p_2, and p_3 are labeled as the same cluster A in the bc_2 column, which means these three patients are undistinguishable under the base clustering bc_2. In other words, either all the patients suffer from heart disease or none of them suffers. However, Fig. 1 shows that patients p_1, p_2, and p_3 distribute at quite different coordinate locations. As a result, we argue that there should exist some hidden or implicit information that can reflect such differences. The grouping structures in other base clusterings (e.g. bc_1, bc_3, bc_4) might be useful to exhibit the hidden distinction among p_1, p_2, p_3 (in the vertical direction of Table I). Furthermore, the neighborhood closeness of patients on their physical conditions could also be applied to explicate the implicit difference among patients (in the horizontal direction of Table I).

This example shows that the IIDness assumption on base clusterings and objects actually causes the aforementioned problems. While the IIDness assumes that each independent object is described by a collection of irrelevant base clustering partitions, there is likely a structural relationship between base clusterings since they are induced from the same data set. How can we discover and describe the coupling relationship between base clusterings? There is also a context surrounding two objects which makes them dependent on each other. How do we design the similarity or distance between objects to capture their relation with other data objects? If there are interactions between both clusterings and objects, how do we integrate such couplings in the clustering ensemble? Those questions suggest a very different assumption for clustering ensembles: non-IIDness [Cao 2014], or more specific, the interdependence based clustering ensemble, which also explicates these inter-dependence relationships in terms of both base clusterings and objects.

Hence, it is our goal to tease out the hidden interdependence relationships as explicit distinguishable attributes in terms of the grouping structures in other base clusterings and the neighborhood closeness of objects, which will be used subsequently to differentiate objects from ambiguous partitions in the clustering ensemble process. In this paper, we introduce the coupled clustering ensemble (i.e. *CCE*) to address the research question of “IIDness”, to uncover the intrinsic coupling relationships between base clusterings, between clusters, and between objects. This paper proposes to incorporate the interactions between base clusterings and between objects, which constitute the behavior interior dimensions. More specifically, *CCE* caters for the cluster label frequency distribution within one base clustering (i.e. intra-coupling of clusterings), the cluster label co-occurrence dependency between distinct base clusterings (i.e. inter-coupling of clusterings), the base clustering aggregation between two objects (i.e. intra-coupling of objects), and the θ-neighborhood relationship among other objects (i.e. inter-coupling of objects), which has been shown to improve the learning accuracy, stability, and robustness. The proposed similarity measures that involve the couplings of base clusterings and objects have been shown to largely tease out the implicit relationships in the data. Substantial experiments have evidenced that the consensus functions incorporated with the interdependent features and behavior interior dimensions significantly outperform those 14 state-of-the-art techniques in terms of the clustering-base, object-based, and cluster-based ensembles, as well as the algorithm to produce base clusterings (i.e. k-means) and some recent clustering algorithms, supported by statistical analysis. This work also verifies that the inter-relationship between objects is essential to the clustering ensemble problem. The coupling of clusterings can enhance the clustering quality in most cases, and the performance gain depends on the quality of base clusterings. The inter-coupling of objects is associated with the consistency of base clustering results, which leads to fluctuating improvement on the clustering quality. The applications in document clustering, as well as on the data sets with large size and dimensionality, further demonstrate the effectiveness, efficiency, and scalability of our proposed *CCE* framework.

The paper is organized as follows. In Section 2, we briefly review the related work. Preliminary definitions are specified in Section 3. Coupling relationships between base clusterings and between objects are specified in Section 4. Section 5 presents the coupled consensus functions together with miscellaneous issues. The effectiveness of *CCE* is shown in Section 6 with intensive experiments. Section 7 discusses the *CCE*. We conclude this work in Section 8.

2. RELATED WORK

We introduce the related work on the clustering ensemble, differentiate the existing consensus function based clustering ensemble from our proposed coupled clustering ensemble.

2.1. Process of Clustering Ensemble

In general, the whole process of the clustering ensemble consists of three steps: building base clusterings, aggregating base clusterings, and post-processing clustering. Various heuristics have been proposed to build the ensemble members, e.g. random initializations, data resampling/subsampling [Kuncheva and Vetrov 2006], random projection and random hyperplane splits [Topchy et al. 2005]. The combination of base clusterings can be constructed by three kinds of method: the consensus functions [Strehl and Ghosh 2002], the categorical clusterings [Gionis et al. 2007], and the direct optimizations [Christou 2011]. The consensus functions focus on the total agreement of all the base clusterings from different perspectives [Li et al. 2010]. The clustering ensemble can also be converted to the problem of clustering categorical data (categorical clustering [Guha et al. 2000; Andritsos et al. 2004], for short) by viewing each attribute as a way of producing a base clustering of the data. However, the direct optimizations [Christou 2011] are substantially performed on the original objective function of clustering rather than exploring the agreement among partial solutions. Finally, the post-processing clustering algorithms are conducted on the consensus building according to the essence of the aggregation structure. For instance, partition-based (e.g., *k-means* [Gionis et al. 2007]) and hierarchy-based (e.g., *single linkage* [Kuncheva and Vetrov 2006]) algorithms are associated with the consensus pairwise matrix, while spectrum-based (e.g., *SPEC* [Fern and Brodley 2004]) and graph-based (e.g., *METIS* [Strehl and Ghosh 2002]) are applicable to the relevant consensus graphs or hypergraphs [Fern and Brodley 2004]. The performance of the clustering ensemble can be greatly enhanced if the algorithms of these three steps are carefully organized.

Here, we focus on building proper consensus functions to aggregate base clusterings, which is the essential element in the clustering ensemble. A consensus function seeks a combination of multiple base clusterings to provide a prior superior input for post-processing clustering. We can construct consensus functions by the following approaches: direct best matching [Li et al. 2010], graph-based mappings [Strehl and Ghosh 2002; Fern and Brodley 2004], statistical mixture models [Topchy et al. 2005], pairwise comparisons [Gionis et al. 2007; Li et al. 2010] and a number of other models. They are all built on the co-associations or pairwise agreements between clusterings (e.g., partition difference *PD* [Li et al. 2010] that focuses on the similarity between partitions and *QMI* [Topchy et al. 2005] that works on the consensus function based on quadratic mutual information), between data objects (e.g., *CSPA*[1] [Strehl and Ghosh 2002]) that induces a similarity measure from partitions and re-clusters objects, or between clusters (e.g., *MCLA* [Strehl and Ghosh 2002] that collapses groups of clusters into meta-clusters and competes for each object to determine the combined clustering). While the clustering ensemble based on consensus functions largely captures the common structure of the base clusterings, and achieves a combined clustering with better quality than individual clusterings, it also faces several issues that have not been explored well in the consensus design. Next, we analyze the problems inherent in the existing work which motivate us to propose the coupled clustering ensemble.

[1] Note that the above categories of approach could overlap; for example, *CSPA* is both a graph-based mapping and pairwise comparison.

2.2. Problems on Consensus Functions

Several papers [Strehl and Ghosh 2002; Gionis et al. 2007] work on the consensus function for clustering ensemble. Heuristics including *CSPA*, *HGPA* based on hypergraph partitioning and *MCLA* [Strehl and Ghosh 2002] solve the ensemble problem by first transforming the base clusterings into a hypergraph representation and then developing consensus functions. Based on *CSPA* and *MCLA*, Fern and Brodley [Fern and Brodley 2004] proposed *HBGF* to consider the similarity between objects and the similarity between clusters collectively. By defining an appropriate distance measure between objects, Gionis et al. [Gionis et al. 2007] mapped the clustering aggregation problem to the weighted correlation clustering problem with linear cost functions. In addition, Topchy et al. [Topchy et al. 2005] introduced a new fusion method *EM* based on a probability model of the consensus partition in the space of contributing clusters and an information-theoretic consensus function *QMI* to effectively combine weak base clusterings. Based on the *EM* model, Nguyen and Caruana [Nguyen and Caruana 2007] presented an *EM*-like consensus algorithm with variations, but they follow the IIDness assumption. Wu et al. [Wu et al. 2013; Wu et al. 2015] provided a systematic study of k-means-based consensus clustering (*KCC* for short) with various utility functions. Most of the existing research on the consensus function has been summarized in [Li et al. 2010], in which the equivalence is revealed between the basic partition difference (*PD*) algorithm and other advanced methods such as Chi-square based approaches.

All of the above methods either fail to address the interactions between base clusterings and between objects (e.g. *CSPA*, *QMI*, *KCC*) or assume independence between them (e.g. *EM*), thus they are IIDness based. Further, the weighted correlation clustering solution proposed in [Gionis et al. 2007] fails to partition the objects if their distance measures are equally 0.5. However, an increasing number of researchers argue that the clustering ensemble is also dependent on the relationship between input partitions [Iam-On et al. 2011; Punera and Ghosh 2007; Domeniconi and Al-Razgan 2009]. Punera and Ghosh [Punera and Ghosh 2007] put forward soft cluster ensemble, in which they used a fuzzy clustering algorithm for the generation of base clusterings. The weighted distance measure [Domeniconi and Al-Razgan 2009] represented a soft relation between a pair of objects and clusters. Unlike our proposed *CCE*, 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. [Iam-On et al. 2011] present a link-based approach to consider the cluster-cluster similarity by connected-triple approach based on the interaction between clusters. The progress of their work and its improved model [Iam-On and Boongoen 2012] is promising, but it overlooks the interaction between objects. So far, no work has been proposed to consider comprehensive couplings, including intra-coupling within and inter-coupling between base clusterings and objects. In this paper, we propose a general and effective model for uncovering the interdependent nature in ensemble clustering.

2.3. Other Related Issues

The clustering ensemble can also be mapped to categorical clustering by treating each base clustering as an attribute [Gionis et al. 2007]. Guha et al. [Guha et al. 2000] proposed *ROCK*, which uses the link-based similarity between two categorical objects. Andritsos et al. [Andritsos et al. 2004] introduced *LIMBO* which is built on the information bottleneck framework for quantifying the relevant information preserved when clustering. In summary, *ROCK* considers the relationship between objects by linkage; *LIMBO* concerns the interaction between different attributes. Neither takes couplings between attributes and between objects into account together, but our proposed *CCE* addresses both.

In our previous work [Wang et al. 2015], we proposed a coupled nominal similarity measure to specify the coupling relationship between attributes. In this paper, we mainly introduce the coupled clustering ensemble, which addresses the problem of seeking the global consensus among base clusterings and also involves the couplings between objects. In addition, the conference version [Wang et al. 2013] of this paper present the coupled clustering ensemble by incorporating the coupling

Table II. An Example of Base Clusterings

C U	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

Table III. List of Main Notations

Variable	Explanation
$\{u_1, \cdots, u_m\}$	The set of m objects U
$\{bc_1, \cdots, bc_L\}$	The set of L base clusterings C
$\{c_j^1, \cdots, c_j^{t_j}\}$	The set of t_j clusters in base clustering bc_j
$\{c_*^1, \cdots, c_*^{t^*}\}$	A final clustering fc^* with t^* clusters
V_j	The set of cluster labels in base clustering bc_j
$v_j^x (\in V_j)$	The cluster label of object u_x in base clustering bc_j
$v_k (\in V_k)$	Any cluster label in base clustering bc_k
δ^{Sim}	The similarity measure
$N_{u_x}^{Sim,\theta}$	The θ-neighbor set of object u_x based on δ^{Sim}
$(BC_j)_{m \times m}$	The associated similarity matrix of objects for bc_j

relationships both between base clusterings and objects. Our proposed models in this paper differ from those in [Wang et al. 2013] mainly on four points: (a) we emphasize the significant aspect of a critical issue: interdependence; (b) we specify the proposed models from the perspective of behavior informatics, introduce the concepts of behavior exterior dimensions and behavior interior dimensions; (c) a concrete toy example on the heart disease data about patients is used to clarify the motivation and contribution of this work; (d) Substantial supporting experiments are added to verify our proposed conclusions from a variety of aspects including relationship discovery, data characteristics analysis, document clustering, applications on data sets with large size and dimensionality, and comprehensive experimental results.

3. PRELIMINARY DEFINITIONS

The problem of the 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 desirable clustering $fc^* = \{c_*^1, \cdots, c_*^{t^*}\}$ with t^* clusters such that the objects inside each cluster c_*^t are close to one another and the objects in different clusters are far from one another.

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 II is a full representation of Table I as an information table consisting of twelve objects (i.e. patients) $\{u_1, u_2, \cdots, u_{12}\}$ 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 $V_4 = \{\alpha, \beta\}$.

Table IV. List of Abbreviations

Abbreviation	Full Name
IaCSC (δ_j^{IaC})	Intra-coupled Clustering Similarity for Clusters
IeRSC ($\delta_{j\|k}$)	Inter-coupled Relative Similarity for Clusters
IeCSC (δ_j^{IeC})	Inter-coupled Clustering Similarity for Clusters
CCSC (δ_j^{C})	Coupled Clustering Similarity for Clusters
IaOSO (δ^{IaO})	Intra-coupled Object Similarity for Objects
IeOSO (δ^{IeO})	Inter-coupled Object Similarity for Objects
CCOSO (δ^{CO})	Coupled Clustering and Object Similarity for Objects
CgC (S_{Cg}^{C})	Proposed Clustering-based Coupling
OC-Ia (S_O^{IaC})	Proposed Intra-coupled Object-based Coupling
OC-H (S_O^{C})	Proposed Hierarchical Object-based Coupling
CrC-Ia ($S_{Cr}^{C} + \delta^{IaO}$)	Proposed Intra-coupled Cluster-based Coupling
CrC-C ($S_{Cr}^{C} + \delta^{CO}$)	Proposed Coupled Cluster-based Coupling

Based on this information-table representation, we use several concepts adapted from our previous work [Wang et al. 2011]. The "set information function" $g_j(v_j^x)$ specifies the set of objects whose cluster labels are 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 $g_4(\alpha)$. Added to this, the "information conditional probability" $P_{k|j}(v_k|v_j^x)$ characterizes the percentage of objects whose cluster labels in base clustering bc_k are v_k among those objects whose cluster labels in base clustering bc_j are exactly v_j^x, formalized as:

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

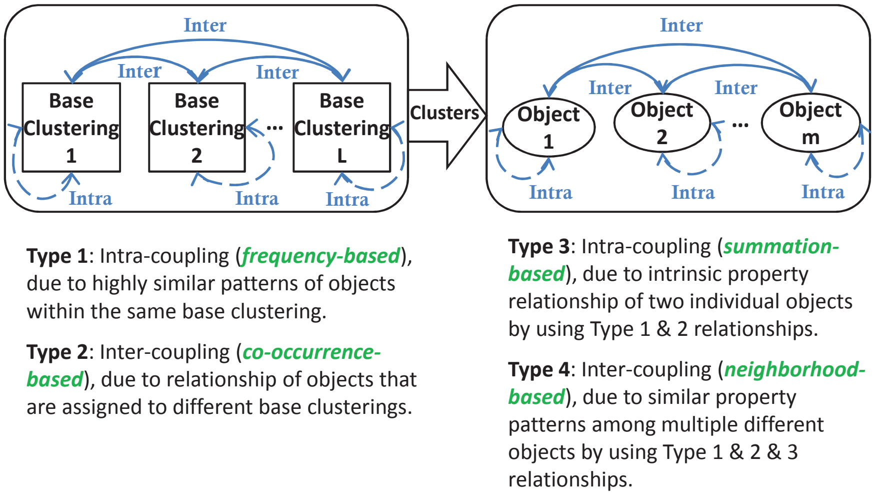
where v_k is a fixed cluster label in base clustering bc_k. Note that $|\cdot|$ is the number of elements in the specific set. For example, we have $P_{2|4}(A|\alpha) = 2/6 = 1/3$.

All these concepts and functions form the foundation of *CCE* for capturing the coupled interactions between base clusterings and between objects. The main notations in this paper are listed in Table III. In addition, several important abbreviations are defined in Table IV to facilitate the reading.

4. COUPLING RELATIONSHIPS

In this section, the coupling relationships in coupled clustering ensemble are proposed in terms of both interactions between base clusterings and between data objects. As described in Fig. 2, the couplings between base clusterings are revealed via the intra-similarity and inter-similarity between cluster labels v_j^x and v_j^y of each base clustering bc_j; and the couplings between objects are specified by defining the intra-similarity and inter-similarity between data objects u_x and u_y.

From the perspective of behavior informatics [Cao et al. 2012], we define "behavior" as the characteristic descriptions of either base clusterings (i.e. attributes) or objects. The behavior is accordingly divided into base clustering behavior and object behavior, which are used to quantify how the base clusterings are allocated and how the objects behave, respectively. "Behavior dimensions" are defined to include "behavior exterior dimensions" and "behavior interior dimensions", which are applied to characterize associated behaviors under different contexts. For the exterior ones, they refer to the explicit information table with multiple given features/variables. For the interior ones, they refer to the implicit coupling relationships/interdependence among base clusterings and among objects. As a result, a converted new information table is to be produced by exploring both the behavior exterior and interior dimensions.



Type 1: Intra-coupling (***frequency-based***), due to highly similar patterns of objects within the same base clustering.

Type 2: Inter-coupling (***co-occurrence-based***), due to relationship of objects that are assigned to different base clusterings.

Type 3: Intra-coupling (***summation-based***), due to intrinsic property relationship of two individual objects by using Type 1 & 2 relationships.

Type 4: Inter-coupling (***neighborhood-based***), due to similar property patterns among multiple different objects by using Type 1 & 2 & 3 relationships.

Fig. 2. Coupling relationships in coupled clustering ensemble, where ←----→ indicates the intra-coupling and ⟷ refers to the inter-coupling.

For example, the consensus among initial results proposed by traditional ensemble strategies works on the behavior exterior dimensions, which is simply presented as the information table shown in Table I. Based on such exterior dimensions, the final partition on patients p_1, p_2, and p_3 is problematic due to the same similarity scores between each pair of them. Accordingly, we try to analyze this data set to tease out the behavior interior dimensions in terms of the grouping structures in other base clusterings and the neighborhood closeness of patients, for instance. They are embodied by the coupling relationships between base clusterings and between objects. The behavior interior dimensions are used to differentiate the objects (e.g. p_1, p_2, and p_3) from ambiguous partitions in Table I, since more hidden/implict information are teased out to help cluster the objects with distinguishable dimensions.

As Fig. 2 indicates, our main task is to model the multiple levels of couplings among base clusterings and among data objects. The intra-coupling of base clusterings (Type 1) is generated by addressing the frequency relationship, while the inter-coupling of base clusterings (Type 2) is induced by considering the co-occurrence relationship. On the other hand, the intra-coupling of objects (Type 3) is quantified by the summation relationship using Type 1 & 2, while the inter-coupling of objects (Type 4) is characterized by the neighborhood relationship using Type 1 & 2 & 3. All of those models, equations, and rationales are to be introduced in details in the following subsections.

4.1. Coupling of Base Clusterings

Since all base clusterings are conducted on the same data objects, intuitively we assume there must be some relationship among these 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 [Wang et al. 2015]. Note that all the component formulae are specified in Section 3.

(1) Intra-coupling of Base Clusterings: We have the Intra-coupled Clustering Similarity for Clusters (*IaCSC* for short) between cluster labels v_j^x and v_j^y of base clustering bc_j defined as

$$\delta_j^{IaC}(v_j^x, v_j^y) = \frac{|g_j(v_j^x)| \cdot |g_j(v_j^y)|}{|g_j(v_j^x)| + |g_j(v_j^y)| + |g_j(v_j^x)| \cdot |g_j(v_j^y)|}, \tag{2}$$

where $g_j(v_j^x)$ and $g_j(v_j^y)$ are set information functions. For example, in Table II, $\delta_j^{IaC}(\alpha, \beta) = 3/4$.

By taking into account the frequency of cluster labels, *IaCSC* characterizes the cluster similarity in terms of cluster label occurrence times. As clarified by [Wang et al. 2015], Equation (2) is a well-defined similarity measure $\delta_j^{IaC} \in [1/3, m/(m+4)]$ 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. Both principles are consistent with the similarity theorem presented in [Lin 1998], in which the commonality corresponds to the product of frequencies and the full description relates to the total sum of individual frequencies and their product. A comparative evaluation on similarity measures for categorical data has been conducted in [Boriah et al. 2008], delivering *OF* and *Lin* as the two best similarity measures among 14 existing measures on 18 data sets. Both these measures assign higher weights to mismatches or matches on frequent values, and the maximum similarity is attained when the attribute values exhibit approximately equal frequencies [Boriah et al. 2008].

Note that the similarity measure proposed for *IaCSC* does not support the maximal self similarity property as widely assumed in traditional works. The reason is that the similarity here is designed to capture the similarity on frequency comparisons rather than the overall proximity. The intra-similarity between cluster label v_j^x and itself may be smaller than that between distinct cluster labels v_j^x and v_j^y when the focus is mainly on the frequency issues. The complete similarity, which is coupled similarity, between attribute values (or cluster labels) also includes the effect of inter-coupled similarity. In other words, the similarity measures are designed to capture the closeness of two cluster labels from different perspectives, and then integrated together to report the overall similarity, in which *IaCSC* delivers just one similarity aspect on the frequency.

Therefore, *IaCSC* considers the interaction between cluster labels within a base clustering bc_j. Intuitively, a larger *IaCSC* similarity score indicates a closer performance of both cluster labels in occurrence frequency pattern. It however 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 next discuss the dependency aggregation, i.e. inter-coupled interaction.

(2) Inter-coupling of Base Clusterings: The Inter-coupled Relative Similarity for Clusters (*IeRSC* for short) between cluster labels v_j^x and v_j^y of base clustering bc_j based on another base clustering bc_k is defined as

$$\delta_{j|k}(v_j^x, v_j^y | V_k) = \sum_{v_k \in \cap} \min\{P_{k|j}(v_k|v_j^x), P_{k|j}(v_k|v_j^y)\}, \tag{3}$$

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 (1). The Inter-coupled Clustering Similarity for Clusters (*IeCSC* for short) between cluster labels v_j^x and v_j^y of base clustering bc_j is defined as

$$\delta_j^{IeC}(v_j^x, v_j^y | \{V_k\}_{k\neq j}) = \sum_{k=1, k\neq j}^{L} \lambda_k \delta_{j|k}(v_j^x, v_j^y | V_k), \tag{4}$$

where $\lambda_k \in [0, 1]$ is the weight for base clustering bc_k, $\sum_{k=1,k\neq j}^{L} \lambda_k = 1$, $V_k (k \neq j)$ 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 [Wang et al. 2015], relative similarity $\delta_{j|k}$ is an improved similarity measure derived from *MVDM* proposed by Cost and Salzberg [Cost and Salzberg 1993]. 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 II, we obtain $\delta_{4|2}(\alpha, \beta|V_2) = 1/3 + 1/2 = 5/6$ and $\delta_4^{IeC}(\alpha, \beta|\{V_1, V_2, V_3\}) = 1/3 \times 5/6 + 1/3 \times 5/6 + 1/3 \times 4/6 = 7/9$ if $\lambda_1 = \lambda_2 = \lambda_3 = 1/3$.

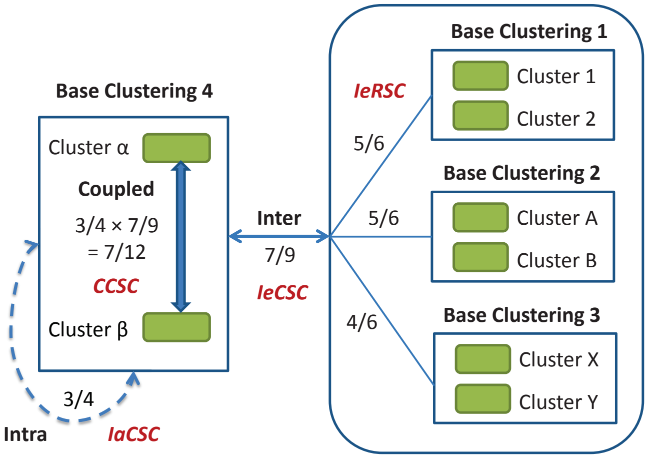


Fig. 3. An example of the coupled similarity for cluster labels α and β, where $\leftarrow\text{- - - -}\rightarrow$ indicates the intra-coupling and $\longleftrightarrow$ refers to the inter-coupling, with the value along each line being the corresponding similarity.

Thus, *IaCSC* captures the base clustering frequency distribution by calculating the occurrence times of cluster labels within one base clustering, and *IeCSC* characterizes the base clustering dependency aggregation by comparing the co-occurrence of the cluster labels in objects among different base clusterings. Intuitively, a greater *IeCSC* similarity value shows a higher probability of both cluster labels in co-occurring consistently with other clustering results. Finally, there is an eligible way to incorporate these two couplings together.

(3) Coupling of Base Clusterings: The Coupled Clustering Similarity for Clusters (*CCSC* for short) between cluster labels v_j^x and v_j^y of clustering bc_j is defined as

$$\delta_j^C(v_j^x, v_j^y|\{V_k\}_{k=1}^L) = \delta_j^{IaC}(v_j^x, v_j^y) \cdot \delta_j^{IeC}(v_j^x, v_j^y|\{V_k\}_{k\neq j}), \tag{5}$$

where δ_j^{IaC} and δ_j^{IeC} are *IaCSC* and *IeCSC*, respectively. As shown in [Wang et al. 2015], $\delta_j^C \in [0, m/(m+4)]$ since $\delta_j^{IaC} \in [1/3, m/(m+4)]$ $(m \geq 2)$ and $\delta_j^{IeC} \in [0, 1]$. In Table II, the coupled similarity between cluster labels α and β is $\delta_4^C(\alpha, \beta|\{V_1, V_2, V_3, V_4\}) = 3/4 \times 7/9 = 7/12$.

As indicated in Equation (5), *CCSC* becomes larger by increasing either *IaCSC* or *IeCSC*. 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 similarity brought by other base clusterings. Intuitively, the multiplication of them indicates the total amount of the cluster closeness; 2) the multiplication method is consistent with the adapted simple matching similarity introduced in [Gan et al. 2007], which considers both the category frequency and matching similarity with 0 or 1.

For example, Fig. 3 summarizes the whole process to calculate the coupled similarity for two cluster labels α and β in Table I. The similarity between α and β is calculated as $7/12$, which is larger than 0 suggested by existing methods. So *CCSC* discloses the implicit relationship for 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. Intuitively, the "intra" here means the calculation of similarity between clusters is limited to only one base clustering, while the "inter" describes how this calculation also considers the involvement of other base clusterings. A higher *CCSC* similarity value demonstrates the larger closeness of two cluster labels with respect to both the frequency in their own clustering result and the co-occurrence with other clustering results.

4.2. Coupling of Objects

In the previous section, we presented the couplings 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 that the objects interact with each other both internally and externally.

(1) Intra-coupling of Objects: 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 u_x and u_y. 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. The Intra-coupled Object Similarity for Objects (*IaOSO* for short) between objects u_x and u_y regarding all the base clustering results of these two objects is defined as

$$\delta^{IaO}(u_x, u_y) = \frac{1}{L} \cdot \sum_{j=1}^{L} \delta_j^C(v_j^x, v_j^y | \{V_k\}_{k=1}^L), \tag{6}$$

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 *CCSC*s 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. Intuitively, the "intra" here represents that the calculation of similarity between objects has nothing to do with other objects. It just involves the two objects to be considered with their internal attributes. The summation here reflects the interdependent relationship via the coupled similarity δ_j^C between relevant attribute values. A greater *IaOSO* similarity score indicates the larger proximity between both objects in terms of their own partition results conducted by different clustering approaches.

For example, the similarity between u_2 and u_3 in Table II is $\delta^{IaO}(u_2, u_3) = 0.655$ and $\delta^{IaO}(u_2, u_{10}) = 0.662$, 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 greater than that between u_2 and u_3, which may prove somewhat misleading in terms of the final clustering in the post-processing stage. To solve this issue, we examine the coupling between objects to further underscore the interaction on the object level.

(2) Inter-coupling of Objects: As indicated in [Guha et al. 2000], the set theory-based similarity measure for categorical values, such as the Jaccard coefficient [Gan et al. 2007], 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 only considers the similarity between the two objects in question, and it is superior to the Jaccard coefficient because it concerns the interactions among base clusterings while the latter is too rough to characterize the pairwise cluster similarity. However, neither *IaOSO* nor Jaccard coefficient reflects the properties of the neighborhood of the objects. Therefore, we are motivated to present our new coupled similarity for objects based on the notions of neighboring and *IeOSO*.

A pair of objects u_x and u_y are defined as θ-neighbors if the following holds

$$\delta^{Sim}(u_x, u_y) \geq \theta, \tag{7}$$

where δ^{Sim} denotes any similarity measure for objects, $\theta \in [0, 1]$ is a given threshold. The θ-neighborhood set of objects u_x can be denoted as

$$N_{u_x}^{Sim,\theta} = \{u_z | \delta^{Sim}(u_x, u_z) \geq \theta\}, \tag{8}$$

which collects all the θ-neighbors of u_x to form an object set $N_{u_x}^{Sim,\theta}$. The similarity measure can be the Jaccard coefficient [Guha et al. 2000] for objects described by categorical attributes, Euclidean dissimilarity [Gan et al. 2007] for objects depicted by continuous attributes, or coupled similarity [Wang et al. 2015] for mixed data. For example, u_3 and u_{10} are the θ-neighbors of object u_2, since $\delta^{Jac}(u_2, u_3) = \delta^{Jac}(u_2, u_{10}) = 1/3 \geq 0.3$ if we adopt the Jaccard co-

Table V. An Example of θ-Neighborhood Domain for Object

Object	θ-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 θ-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}	$\{u_3, u_6, u_7, u_8, u_{11}, u_{12}\}$

efficient as the similarity measure and set $\theta = 0.3$, and then the θ-neighborhood set of u_2 is $N_{u_2}^{Jac,0.3} = \{u_1, u_3, u_4, u_5, u_6, u_7, u_{10}, u_{11}\}$.

Further, we can embody the inter-coupled interaction between different objects by exploring the relationship between their θ-neighborhoods. Intuitively, objects u_x and u_y more likely belong to the same cluster if they have a larger overlap in their θ-neighborhood sets $N_{u_x}^{Sim,\theta}$ and $N_{u_y}^{Sim,\theta}$. Thus, we use the common θ-neighbors to define the inter-coupled similarity for objects. The inter-coupled Object Similarity for Objects (*IeOSO* for short) 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, based on similarity δ^{Sim} as below

$$\delta^{IeO}(u_x, u_y|U, \delta^{Sim}, \theta) = \frac{1}{m} \cdot |\{u_z \in U | u_z \in N_{u_x}^{Sim,\theta} \cap N_{u_y}^{Sim,\theta}\}|, \tag{9}$$

where $N_{u_x}^{Sim,\theta}$ and $N_{u_y}^{Sim,\theta}$ are the θ-neighborhood sets of objects u_x and u_y based on similarity measure δ^{Sim}, respectively. For example, $\delta^{IeO}(u_2, u_3|U, \delta^{Sim}, \theta) = 0.583$ and $\delta^{IeO}(u_2, u_{10}|U, \delta^{Sim}, \theta) = 0.417$ when setting δ^{Sim} to be the Jaccard coefficient and $\theta = 0.3$.

Therefore, *IeOSO* builds the inter-coupling relationship between each pair of objects by capturing the global knowledge of their θ-neighborhood. Intuitively, the "inter" specifies that the calculation of similarity between objects also concerns other objects if they are in a θ-neighborhood relationship. A higher *IeOSO* similarity value exhibits the closer collections of neighbors for both objects.

(3) Coupling of Objects: The intra-coupling and inter-coupled interactions are considered together to induce the coupled similarity for objects by exactly specifying the similarity measure δ^{Sim} to be *IaOSO* as defined in Equation (6). The Coupled Clustering and Object Similarity for Objects (*CCOSO* for short) between objects u_x and u_y is defined when δ^{Sim} is regarded as δ^{IaO}

$$\begin{aligned} \delta^{CO}(u_x, u_y|U, \theta) &= \delta^{IeO}(u_x, u_y|U, \delta^{IaO}, \theta) \\ &= \frac{1}{m} \cdot |\{u_z \in U | u_z \in N_{u_x}^{IaO,\theta} \cap N_{u_y}^{IaO,\theta}\}|, \end{aligned} \tag{10}$$

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-coupling 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 this the coupled clustering and object similarity for objects. Intuitively, a greater *CCOSO* score shows the larger similarity between objects in terms of their own clustering results and the neighborhood relationships with other objects.

For example, the corresponding θ-neighbors of objects u_2, u_3 and u_{10} are described in Table V, here $\theta = 0.65$. From this table, we observe that the number of common θ-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 solves the uncertain assignment problem raised in Section 1. Based on the equation of *CCOSO*, we obtain $\delta^{CO}(u_2, u_3|U, \theta) = 0.75$ and $\delta^{CO}(u_2, u_{10}|U, \theta) = 0.5$. This 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 $\delta^{IaO}(u_2, u_3) < \delta^{IaO}(u_2, u_{10})$.

4.3. Behavior-based Explanations

From the perspective of behavior informatics, as mentioned in Section 1, behavior refers to the characteristic descriptions of either base clusterings (i.e. attributes) or objects, which can be further divided into the base clustering behavior and object behavior. They are used to measure how the base clusterings are allocated and how the objects behave, respectively.

On one hand, the concept on the coupled clustering similarity for clusters explicates the behavior interior dimensions for base clusterings. For each base clustering bc_j, we may consider to include a matrix M_j to display the coupled similarity for each pair of clusters obtained in this base clustering. If there are t_j clusters in the base clustering bc_j, then the matrix M_j has $t_j \times t_j$ entries in which the (x, y) entry quantifies the coupled similarity between clusters c_j^x and c_j^y in bc_j. Formally, we have:

$$M_j(x, y) = \delta_j^C(v_j^x, v_j^y | \{V_k\}_{k=1}^L). \tag{11}$$

Accordingly, we obtain L such matrices $M_1, M_2, \cdots, M_L$ to show the behavior interior dimensions for base clusterings. Those matrices reveal the interdependent relationships among base clusterings via the co-occurrence of objects in the same clusters. Based on such interior dimensions for base clusterings, we have teased out the implicit knowledge hidden in the naive information table (i.e. behavior exterior dimensions), as an example shown in Table II.

On the other hand, both concepts on the intra-coupled object similarity for objects and the coupled clustering and object similarity for objects exhibit the behavior interior dimensions for objects. For each object u_x, we may add a row vector U_x^{IaO} or U_x^{CO} at the end of each row in the information table to show the intra-coupled similarity or coupled similarity for every pair of objects. The sizes of row vectors U_x^{IaO} and U_x^{CO} are both $1 \times m$ since there are in total m objects, in which the yth components correspond to the respective intra-coupled similarity and coupled similarity between objects u_x and u_y. Formally, we have:

$$U_x^{IaO}(y) = \delta^{IaO}(u_x, u_y),\ U_x^{CO}(y) = \delta^{CO}(u_x, u_y | U, \theta) \tag{12}$$

Therefore, we obtain m row vectors $U_1^{IaO}, U_2^{IaO}, \cdots, U_m^{IaO}$ and another m row vectors U_1^{CO}, $U_2^{CO}, \cdots, U_m^{CO}$ to represent the behavior interior dimensions for objects. Those vectors disclose the interdependent relationships among objects via the θ-neighborhood connections.

The behavior interior dimensions are extracted from the behavior exterior dimensions, as shown in Fig. 4. The behavior exterior dimensions are directly reflected in the information table with m objects and L base clusterings. The behavior interior dimensions are teased out for base clustering and for objects individually. The former is presented as a set of L similarity matrices $\{M_j\}_{j=1}^L$ to quantify the pairwise similarity between clusters in each base clustering, while the latter one is exhibited as two sets of m similarity vectors $\{U_i^{IaO}\}_{i=1}^m$ and $\{U_i^{CO}\}_{i=1}^m$ to measure the intra-coupled similarity and coupled similarity between objects, respectively. By using such behavior interior dimensions together with the behavior exterior dimensions, we are able to perform the clustering ensemble effectively, since more information is exposed to be available to distinguish the differences among data. Below, we make use of all the components including $\{M_j\}_{j=1}^L$, $\{U_i^{IaO}\}_{i=1}^m$, $\{U_i^{CO}\}_{i=1}^m$ in Fig. 4 to build the coupled consensus function.

For example in Table II, for M_4, we have calculated $M_4(2, 1) = \delta_4^C(\alpha, \beta | \{V_1, V_2, V_3, V_4\})) = M_4(1, 2) = \delta_4^C(\beta, \alpha | \{V_k\}_{k=1}^4) = 7/12 \approx 0.583$ as illustrated in Fig. 3 in Section 4.1. For the other entries of matrix M_j, we follow the same way as defined in Equations (2), (4), and (5), thus have the behavior interior dimensions displayed as follows. The behavior interior dimensions for base clusterings consist of four matrices M_1, M_2, M_3 and M_4:

$$M_1 = \begin{pmatrix} 0.714 & 0.638 \\ 0.638 & 0.778 \end{pmatrix},\ M_2 = \begin{pmatrix} 0.714 & 0.667 \\ 0.667 & 0.778 \end{pmatrix}$$

$$M_3 = \begin{pmatrix} 0.667 & 0.546 \\ 0.546 & 0.800 \end{pmatrix},\ M_4 = \begin{pmatrix} 0.750 & 0.583 \\ 0.583 & 0.750 \end{pmatrix}$$

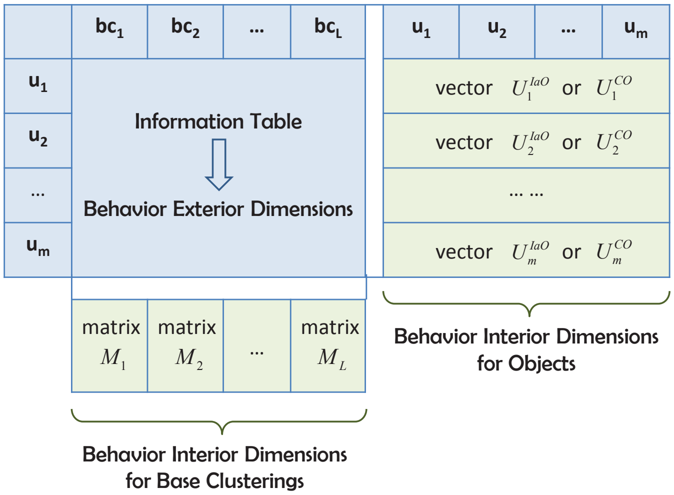

bc$_1$
bc$_2$
...
bc$_L$
u$_1$
u$_2$
...
u$_m$
Information Table
Behavior Exterior Dimensions
vector U_1^{IaO} or U_1^{CO}
vector U_2^{IaO} or U_2^{CO}
... ...
vector U_m^{IaO} or U_m^{CO}
matrix M_1
matrix M_2
matrix M_L
Behavior Interior Dimensions for Objects
Behavior Interior Dimensions for Base Clusterings

Fig. 4. Behavior interior dimensions for base clusterings and objects.

Each entry in matrix M_j exhibits the pairwise coupled similarity between clusters in every base clustering. For the three patients p_1, p_2, p_3 in Table I corresponding to u_2, u_3, u_{10} in Table II, the intra-coupled similarity scores between objects u_2 and u_3 and between objects u_2 and u_{10} are $\delta^{IaO}(u_2, u_3) = 0.655$ and $\delta^{IaO}(u_2, u_{10}) = 0.662$, respectively, as shown in Section 4.2. For the other components of vector U_i^{IaO}, we apply the Equation (6), and have the intra-coupled object similarity based behavior interior dimensions U_i^{IaO} for objects as below:

$$
\begin{aligned}
U_2^{IaO} &= (0.674\ \ 0.727\ \ 0.655\ \ 0.674\ \ 0.651\ \ 0.655\ \ 0.685\ \ 0.650\ \ 0.608\ \ 0.662\ \ 0.685\ \ 0.650)\\
U_3^{IaO} &= (0.685\ \ 0.655\ \ 0.761\ \ 0.685\ \ 0.662\ \ 0.761\ \ 0.707\ \ 0.672\ \ 0.714\ \ 0.684\ \ 0.707\ \ 0.672)\\
U_{10}^{IaO} &= (0.608\ \ 0.662\ \ 0.684\ \ 0.608\ \ 0.639\ \ 0.684\ \ 0.714\ \ 0.733\ \ 0.691\ \ 0.745\ \ 0.714\ \ 0.733)
\end{aligned}
$$

Each component in vector U_i^{IaO} reflects the pairwise intra-coupled similarity between objects. With respect to the coupled clustering and object similarity, we have the coupled similarity scores between objects u_2 and u_3 and between u_2 and u_{10} calculate to be $\delta^{CO}(u_2, u_3|U, \theta) = 0.75$ and $\delta^{CO}(u_2, u_{10}|U, \theta) = 0.5$ when $\theta = 0.65$ based on Table V, as indicated in Section 4.2. Likewise, the associated behavior interior dimensions U_i^{CO} for objects when $\theta = 0.65$ are accordingly obtained by applying Equation (10) as follows:

$$
\begin{aligned}
U_2^{CO} &= (0.500\ \ 0.727\ \ 0.750\ \ 0.500\ \ 0.333\ \ 0.750\ \ 0.667\ \ 0.500\ \ 0.833\ \ 0.500\ \ 0.667\ \ 0.500)\\
U_3^{CO} &= (0.583\ \ 0.750\ \ 0.761\ \ 0.583\ \ 0.417\ \ 0.833\ \ 0.750\ \ 0.583\ \ 0.750\ \ 0.583\ \ 0.750\ \ 0.583)\\
U_{10}^{CO} &= (0.500\ \ 0.500\ \ 0.583\ \ 0.500\ \ 0.333\ \ 0.583\ \ 0.583\ \ 0.583\ \ 0.500\ \ 0.745\ \ 0.583\ \ 0.583)
\end{aligned}
$$

Each component in U_i^{CO} quantifies the pairwise coupled similarity between objects.

By contrast, for the current clustering ensemble models such as *CSPA*, *HGPA*, and *MCLA* [Gionis et al. 2007] as well as *EM* and *QMI* [Topchy et al. 2005], the behavior interior dimensions are the same as the behavior exterior dimensions. In other words, those approaches are performed directly on the behavior exterior dimensions represented by an information table. Accordingly, the green area in Fig. 4 for them is empty: they have neither behavior interior dimensions for base clusterings nor behavior interior dimensions for objects. For some advanced and more recent methods, including weighted distance model [Domeniconi and Al-Razgan 2009] and link-based model [Iam-On et al. 2011; Iam-On and Boongoen 2012], they only deliver partial behavior interior dimensions (either for base clusterings or for objects) in an implicit way. Different from those existing methods, our proposed *CCE* has been provided with formalized definitions/formulae and underpinning theoretical support for the construction and quantification of complete behavior interior dimensions.

5. CONSENSUS FUNCTION MODELS

There are many different ways to define the consensus function such as pairwise agreements between base clusterings, co-associations between data objects, and interactions between clusters. To the best of our knowledge, some of the criteria focus on the estimation of similarity between base clusterings [Li et al. 2010; Topchy et al. 2005], some are based on the similarity between data objects [Strehl and Ghosh 2002], and others are associated with the similarity between clusters [Fern and Brodley 2004; Iam-On et al. 2011]. In the following, we specify the coupled models of clustering-based, object-based, and cluster-based criteria individually by applying the various interdependence relationships proposed in Section 4.

5.1. Traditional Consensus Function

Firstly, we summarize three categories of the most popular consensus functions that are in current use for clustering ensembles: clustering-based model, object-based model, and cluster-based model.

(1) Clustering-based Model: 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 the 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:

$$BC_j(x,y) = \begin{cases} 1 & \text{if } v_j^x = v_j^y, \\ 0 & \text{otherwise}, \end{cases} \tag{13}$$

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*) [Li et al. 2010]:

$$S_{Cg}(bc_{j_1}, bc_{j_2}) = \sum_{1\le x,y\le m} \left[BC_{j_1}(x,y) - BC_{j_2}(x,y)\right]^2, \tag{14}$$

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

(2) Object-based Model: 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, an overall similarity matrix BC_t^* with a finer resolution is produced [Strehl and Ghosh 2002]. Formally, we have:

$$BC_t^*(x,y) = \frac{1}{L}\cdot\sum_{j=1}^{L} BC_j(x,y). \tag{15}$$

The entry of the induced overall similarity matrix BC_t^* 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 inadequate since only 1 and 0 are considered as defined in Equation (13), which is issue (i). The relationship that is neither within nor between base clusterings (i.e., bc_{j_1} and bc_{j_2}) is explicated, i.e. issue (ii). In addition, there is issue (iii): most existing methods [Gionis et al. 2007; Christou 2011] only use the similarity measure between objects when clustering them, which thus does not involve the context (i.e. θ-neighborhood).

(3) Cluster-based Model: 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 [Strehl and Ghosh 2002]. The idea is to yield object-wise confidence estimates of cluster membership, to group and then to collapse related clusters represented as hyper-edges. The similarity measure of clusters in *MCLA* is the Jaccard matching coefficient [Gan et al. 2007]:

$$S_{Cr}(c_{j_1}^{t_1}, c_{j_2}^{t_2}) = \frac{|c_{j_1}^{t_1} \cap c_{j_2}^{t_2}|}{|c_{j_1}^{t_1} \cup c_{j_2}^{t_2}|}, \tag{16}$$

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.

5.2. Coupled Consensus Function

Based on the above traditional consensus functions, we propose the coupled models by addressing the coupling relationships between base clusterings and between objects.

(1) Coupled Clustering-based Model

Regarding Equations (13) and (14), alternatively, we can focus on the entry $BC_j(x, y)$ to incorporate the coupling of base clusterings as follows:

$$BC_j^C(x, y) = \delta_j^C(v_j^x, v_j^y | \{V_k\}_{k=1}^L), \tag{17}$$

$$S_{Cg}^C(bc_{j_1}, bc_{j_2}) = \sum_{1 \leq x, y \leq m} [BC_{j_1}^C(x, y) - BC_{j_2}^C(x, y)]^2, \tag{18}$$

where δ_j^C refers to *CCSC* in Equation (5). 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:

$$fc^* = \arg_{c^1, \cdots, c^{t^*}} \min \sum_{j=1}^{L} S_{Cg}^C(fc, bc_j), \tag{19}$$

where $fc = \{c^1, \cdots, c^{t^*}\}$ denotes the candidate set of clusters for final clustering fc^*. According to [Topchy et al. 2005], the optimization problem in Formula (19) then can be heuristically approached by *k-means* operating in the normalized object-label space OL with each entry to be:

$$OL(u_x, v_j^y) = \delta_j^C(v_j^x, v_j^y | \{V_k\}_{k=1}^L) - \mu^y(\delta_j^C), \tag{20}$$

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 $\delta_j^C(v_j, v_j^y | \{V_k\}_{k=1}^L)$ for the cluster label v_j^y with all possible attribute values $v_j \in V_j$.

Therefore, the clustering-based coupling addresses the intra-coupling and inter-coupling of base clusterings to form the coupled consensus function *CgC*.

(2) Coupled Object-based Model

To solve the issue (i) and issue (ii) raised on Equation (15), we regard the entry $BC_{Ia}^*(x, y)$ of the overall similarity matrix to be *IaOSO*:

$$S_O^{IaC}(u_x, u_y) = BC_{Ia}^*(x, y) = \delta^{IaO}(u_x, u_y), \tag{21}$$

where δ^{IaO} is defined in Equation (6). 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 defined in Equation (5), in the same way as Equation (17); then, the overall similarity matrix BC_{Ia}^* is obtained by averaging the associated similarity matrix BC_j over all the base clusterings according to Equation (15).

Afterwards, *METIS* is applied to the overall similarity matrix BC_{Ia}^* to produce the final clustering fc^*. We denote this newly proposed intra-coupled object-based coupling method as *OC-Ia*.

Further considering the issue (iii) for Equation (15), both the intra-couplings and inter-couplings of clusterings and of objects are incorporated as follows:

$$S_O^C(u_x, u_y) = BC_c^*(x, y) = \delta^{CO}(u_x, u_y|U, \theta), \tag{22}$$

where δ^{CO} is defined in Equation (10). We would like to maximize the sum of $\delta^{CO}(u_x, u_y|U, \theta)$ for data object pairs u_x, u_y belonging to a single cluster, and at the same time minimize the sum of $\delta^{CO}(u_x, u_y|U, \theta)$ 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:

$$fc^* = \underset{c^1, \cdots, c^{t^*}}{\arg} \max \sum_{t=1}^{t^*} m_t \cdot \sum_{u_x, u_y \in c^t} \frac{S_O^C(u_x, u_y) \cdot m}{m_t^{1+2f(\theta)}}, \tag{23}$$

where c^t denotes the tth cluster of size m_t, m is the total number of objects, and $f(\theta) = (1-\theta)/(1+\theta)$, θ is the threshold defined in neighborhood. The rationale of the above function is twofold: on one hand, one of our goals is to maximize $\delta^{CO}(u_x, u_y|U, \theta)$ 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 $m_t^{1+2f(\theta)}/m$ [Guha et al. 2000]; 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 a clustering in which all objects are assigned to a single cluster, and avoids objects with very small coupled similarity value between them from being put in the same cluster [Guha et al. 2000]. Next, we adapt the standard agglomerative hierarchical clustering algorithm to obtain the final clustering fc^* by solving Equation (23) [Guha et al. 2000]. We abbreviate this newly proposed hierarchical object-based coupling to *OC-H*.

The intra-coupled object-based coupling examines the intra-coupling and inter-coupling of base clusterings as well as the intra-coupling of objects to form the coupled consensus function *OC-Ia*, while the hierarchical object-based coupling considers both the intra-coupling and inter-coupling of base clustering and objects to build the coupled consensus *OC-H*.

(3) Coupled Cluster-based Model

The similarity measure S_{Cr} introduced in Equation (16) considers neither coupling between base clusterings nor interaction between objects. Therefore, it lacks the capability to reflect the essential link and relationship among data. 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 coupling 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:

$$S_{Cr}^C(c_{j_1}^{t_1}, c_{j_2}^{t_2}) = \frac{1}{m_{t_1} m_{t_2}} \sum_{u_x \in c_{j_1}^{t_1}, u_y \in c_{j_2}^{t_2}} S_O(u_x, u_y), \tag{24}$$

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, and can be either δ^{IaO} defined in Equation (6) or δ^{CO} defined in Equation (10). 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 [Strehl and Ghosh 2002]. We denote the cluster-based coupling as *CrC* (including *CrC-Ia* with δ^{IaO} and *CrC-C* with δ^{CO}).

The intra-coupled cluster-based coupling considers the intra-coupling and inter-coupling of base clusterings together with the intra-coupling of objects to define the coupled consensus function *CrC-*

Ia, while the coupled object-based coupling concerns both the intra-coupling and inter-coupling of base clustering and objects to construct the coupled consensus function *CrC-C*.

Note that the difference between objective functions defined in Equation (19) and Equation (23) is that the former only considers the coupling relationship between base clusterings but the latter also addresses the interdependence between objects. The inclusion of interactions between base clusterings and between objects is dependent on the accuracy and consistency of base clusterings, which is empirically studied in Section 6.4.

5.3. Discussions

Exploring the relationships among base clusterings and among objects is important because it gives more accurate similarity findings among base clusterings and among objects, thus improving the quality of final clustering results. However, due to the rapid increase in the data volume (in terms of the number of objects and also the number of attributes associated with an object), there are at last two main issues existing for our proposed coupled ensemble models: Big Data Issue and Computational Complexity Issue. In this section, we will discuss each of them in details.

(1) Big Data Issue: In case the data set is large, random sampling and labeling enable the pairwise similarity-based *CCE* to reduce the number of objects to be considered, and ensure that the input data set fits the main memory. Efficient algorithms for selecting random samples can be found in [Guha et al. 2000]. As indicated in [Gionis et al. 2007], sampling $O(\log m)$ objects is sufficient to guarantee that at least one object in a large cluster will be selected with high probability. Afterwards, *CCE* assigns the remaining data objects to the clusters generated by the sampled objects, according to the similarity between each object and a fraction of objects from every cluster. If the sum of similarity between the object u_x to be labeled and the objects chosen from a final cluster c_*^t is maximum, then the object u_x is allocated to the tth final cluster c_*^t.

(2) Computational Complexity Issue: We have discussed the coupled clustering ensemble from the perspectives of coupling of clusterings, coupling of objects, and coupled consensus functions. The computational complexity to calculate the coupled clustering similarity for clusters *CCSC* is $O(LT^3)$, and the computational complexity to calculate the coupled clustering and object similarity for objects *CCOSO* is $O(L^2T^3+2m)$, where L is the number of base clusterings, T is the maximum number of clusters in all the base clusterings, and m is the total number of objects. The detailed pseudocodes are specified in the conference version [Wang et al. 2013] of this paper.

6. EMPIRICAL STUDY

This section presents the performance of *CCE* with the clustering-based (*CgC*), object-based (*OC-Ia* and *OC-H*), and cluster-based (*CrC-Ia* and *CrC-C*) couplings. Experiments are performed on synthetic and real-life data sets to discover the implicit relationships between base clusterings and between objects, to validate accuracy, stability, and robustness of various consensus functions, and to explore the dependency between data characteristics and clustering quality. In addition, we propose the following assumptions for all the groups of experiments in this section.

How to Establish the Number of Final Clusters: The automatic identification of the appropriate number of clusters is a deep research problem that has attracted significant attention [Gionis et al. 2007; Kuncheva and Vetrov 2006; Wang et al. 2010]. There are four ways to handle this issue: imposing a hard constraint on the number of clusters or on their quality, model selection, finding the size t^* of final clustering by similarity analysis, and nonparametric estimation. In our experiments, for simplicity, the number of clusters t^* is fixed, the same as the number of classes in each data set. The different ways to determine t^* can also be incorporated into our proposed coupled consensus functions.

How to Generate Base Clusterings: There are several methods of providing diverse base clusterings: using different clustering algorithms [Gionis et al. 2007], employing random or different parameters for some algorithms [Iam-On et al. 2011], and adopting random sub-sampling or random projection of the data [Fern and Brodley 2004]. Since our focus is mainly on the consensus function, we use *k-means* on random sub-sampling [Fern and Brodley 2004] of the data as the base

Table VI. Description of Data Sets

Data Set	m	n	t^p	Source
Sy1	200	2	2	modified
Sy2	400	6	4	modified
Iris	150	4	3	UCI
Wine	178	13	3	UCI
Seg	210	19	7	UCI
Glass	214	9	6	UCI
Ecoli	336	7	8	UCI
Ionos	351	34	2	UCI
Blood	748	5	2	UCI
Vowel	990	10	11	UCI
Yeast	1,484	8	10	UCI

clustering algorithm in our experiments. The number of base clusterings is pre-defined for each data and remains the same in all runs.

How to Post-process Clustering: In the proposed *CCE*, 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 re-cluster the objects using any reasonable similarity-based clustering algorithm. In our experiments, we choose *k-means*, *agglomerative algorithm* [Guha et al. 2000] and *METIS* [Strehl and Ghosh 2002] due to their popularity in the clustering ensemble.

6.1. Data Sets

The experimental evaluation is conducted on 11 data sets, including two synthetic data sets (i.e., Sy1 and Sy2, which are 2-Gaussian modified from [Strehl and Ghosh 2002] and 4-GaussianN modified from [Kuncheva and Vetrov 2006], respectively) and nine real-life data sets from UCI [Frank and Asuncion 2010]. Table VI 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 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 is initially created to follow the two-dimension Gaussian distribution and then added with four more dimensions of uniform random noise in the way presented in [Kuncheva and Vetrov 2006]. There are another four document data sets and eight data sets with larger size and dimensionality, which are specified in Section 6.5 and Section 6.6, respectively.

6.2. Selection of Baseline Approaches and Parameters

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

(1) Clustering-based: Besides the partition difference (*PD*) proposed in [Li et al. 2010], *QMI* is also an effective clustering-based criterion [Topchy et al. 2005], which has proved to be equivalent to *Category Utility Function* in [Li et al. 2010]. K-means-based consensus clustering (*KCC*) with the default choice of utility function NU_H is also used to make comparisons. We will compare the clustering-based coupling (*CgC*) with its baseline method *PD* [Li et al. 2010], *EM* and *QMI* [Topchy et al. 2005], as well as *KCC* [Wu et al. 2015].

(2) Object-based: In this group, we will compare the intra-coupled object-based coupling *OC-Ia* with its baseline method *CSPA* [Strehl and Ghosh 2002], and compare the hierarchical object-based coupling *OC-H* with *CSPA* [Strehl and Ghosh 2002] and with the categorical clustering algorithms: *ROCK* [Guha et al. 2000] (the baseline method of *OC-H*) and *LIMBO* [Andritsos et al. 2004].

(3) Cluster-based: Based on *MCLA* [Strehl and Ghosh 2002], *HBGF* is another promising cluster-based criterion [Fern and Brodley 2004]. It also collectively considers the similarity between objects and clusters but lacks the discovery of coupling. Iam-On et al. [Iam-On et al. 2011] 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 methods *MCLA* [Strehl and Ghosh 2002], *HBGF* [Fern and Brodley 2004], and *LB* [Iam-On et al. 2011; Iam-On and Boongoen 2012] (including *LB-P* and *LB-S*[2]).

As indicated at the beginning of this section, *k-means* on random sub-sampling [Fern and Brodley 2004] 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*. Here, *OC-H* is built based on *ROCK* [Guha et al. 2000], thus *agglomerative algorithm* is adopted to do the post-processing as *ROCK* does. But *METIS* is much more efficient than *agglomerative algorithm*, so we use *METIS* to post-process *OC-Ia*. The following parameters of the clustering ensemble are especially important:

- θ: The θ-neighbor threshold is defined to be the average *IaOCO* and Jaccard coefficient [Guha et al. 2000] 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 $L = 10$. The reason for selecting $L = 10$ will be explained in Section 6.3.
- 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 of *IeCSC* is simplified as $\lambda_k = 1/(L-1) = 1/9$.
- NR: The number of runs for each clustering ensemble is fixed as $NR = 50$ to obtain corresponding average results for the evaluation measures below.

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

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) [Cai et al. 2005] and normalized mutual information (NMI) [Cai et al. 2005]. Note that the external information about class labels is available for the data sets in Table VI, so we apply the external clustering validation measures such as AC and NMI in this paper, rather than the internal clustering validation measures introduced in [Liu et al. 2013]. We also judge the stability of multiple runs by using the combined stability index (CSI) [Kuncheva and Vetrov 2006], as well as the robustness [Topchy et al. 2005] of the clustering ensemble by comparing the average AC, NMI, and CSI scores across different data sets. In brief, AC and NMI describe the degree of approximation between the obtained clusters and the true classes. CSI reveals the stability between them across $NR = 50$ runs, and reflects the deviation of the results across different runs. The larger the AC or NMI or CSI is, the better the clustering ensemble algorithm is. Note that the correspondence problem on mapping between the derived clusters and the known classes needs to be solved before evaluation. The optimal correspondence can be obtained using the Hungarian method [Topchy et al. 2005] with $O((t^p)^3)$ complexity.

6.3. Experimental Results

Based on the evaluation measures (i.e., AC, NMI and CSI), Table VII 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. Below, we report the experimental results on implicit relationship

[2] The performance of the model (i.e., *WTU+SPEC*) proposed in [Iam-On and Boongoen 2012] is between that of *LB-P* (i.e., *CSM+PAM* [Iam-On et al. 2011]) and that of *LB-S* (i.e., *CSM+SPEC* [Iam-On et al. 2011]), thus we only report the results of *LB-P* and *LB-S*.

Table VII. 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

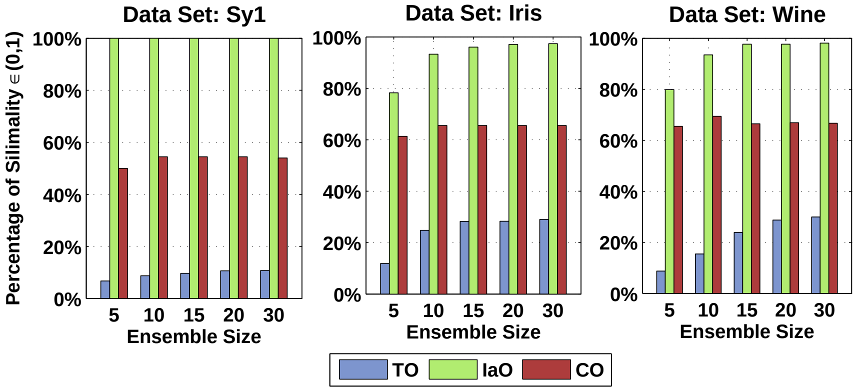


Fig. 5. Percentage of pairwise similarity $\in (0, 1)$ between objects among all the similarity values.

discovery, clustering-based comparison, object-based comparison, and cluster-based comparison individually.

(1) Implicit Relationship Discovery

Prior to the experiments on the clustering ensemble, we first compare the implicit relationship revealed by different similarity measures between objects. The similarity measures to be compared are listed here: traditional measure (*TO* for short) specified in Formulae (13) and (15), intra-coupled object similarity (*IaO* for short) defined in Equation (6), coupled clustering and object similarity (*CO* for short) defined in Equation (10). The unknown relationship captured by similarity measures is quantified as the percentage of similarity values that fall within the open interval $(0, 1)$ among all of the pairwise similarities between objects. Note that the ratios reported in this comparison are the average values across 50 runs of generating base clusterings.

Fig. 5 presents the percentage of similarity values $\in (0, 1)$ (axis y) for three similarity measures (i.e., *TO*, *IaO* and *CO*) on three data sets (i.e., Sy1, Iris, and Wine) with different ensemble sizes L ranging from 5 to 30 (axis x). It is remarkable to note that the proportions of similarity values $\in (0, 1)$ for *IaO* and *CO* are much higher than for *TO*. This empirical evidence signifies that our proposed *IaO* and *CO* are capable of discovering the hidden relationships among data objects, while the *TO* performs rather poorly by mostly assigning 0 and 1 to the similarity between objects.

Another interesting observation is that the percentage score of *IaO* is larger than that of *CO*. This is probably due to the fact that the similarity measure *CO* is filtered and refined from *IaO*, which means *CO* may amplify several *IaO* similarity values and also diminish some *IaO* values according to the neighborhood coupling. In essence, IaO captures a partial picture of the similarity between objects, while CO provides a global view in terms of the context around objects.

A third discovery is that the ensemble size $L = 10$ is large enough to capture the relationship between data objects, as compared to $L = 15, 20, 30$. It can be also observed that the percentages of TO and IaO have an increasing trend when L goes up, while the ratio of CO keeps fluctuating. The reason is that the likelihood that the TO and IaO values will take 0 become smaller with the increasing number of base clusterings. However, the opportunity for CO to be evaluated as 0 is uncertain since the average threshold for defining θ-neighbors in Equation (7) also increases, which probably leads to a smaller set of θ-neighbors. The ensemble size can influence the final consensual result, however, we mainly focus on how the ensemble is generated rather than how the ensemble size matters. So the number of base clusterings is kept fixed in all the groups of experiments below to test how our proposed ensemble strategies work. According to Fig. 5, we select $L = 10$ to preserve the ability to discover sufficient relationships but with relatively low computational complexity, which is also applied in many other papers such as [Iam-On et al. 2011].

In the following sections, 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 joint effect of couplings between base clusterings and between objects. We analyze the clustering performance individually by considering the couplings step by step within each group of experiments, although a comparison across these three groups is beyond the 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). The 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.

(2) Clustering-based Comparison

Fig. 6 shows the performance comparison of different clustering-based ensemble methods over two synthetic and nine real-life data sets in terms of AC, NMI and CSI. 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*, *QMI*, and *KCC*. Specifically, in terms of accuracy, the AC improvement rate ranges from 1.37% (*KCC* on Sy2) to 12.71% (*EM* on Vowel) only except on Ecoli (i.e. *KCC* performs slightly better than *CgC* on this data), and there has been significant CSI improvement (from 0.61% to 49.83%) except in one case: Glass. Overall, the average improvement rate of *CgC* on AC across all the other methods over all the data sets is 3.42%, and the average improvement rate of *CgC* on CSI is 7.16%. Also, in several data sets such as Sy1, Sy2, Wine, Seg, Ionos, Blood, Vowel and Yeast, the AC measures exceed the maximum of AC in the corresponding base clusterings, i.e. Max(AC) in Table VII. All the AC and CSI values of *CgC* are higher than the corresponding average values of base clustering. Another observation is that for the compared consensus functions, *KCC* is the best in most cases, followed by *QMI* and *PD* with *EM* being the worst. However, our proposed *CgC* outperforms all the algorithms compared on almost every data set. A similar situation can also be observed when NMI is used to evaluate clustering quality. Statistical analysis, namely the t-test, has been carried out on the AC and NMI of our *CgC*, at a 95% significance level. The null hypothesis that *CgC* is better than base clusterings and the best result of other methods in terms of AC and NMI is accepted.

In addition, it seems that the improvement level of *CgC* upon other methods is associated with the quality of base clusterings: the better quality of base clusterings corresponds to a relatively smaller level of improvement. This point of view will be justified later in Section 6.4.

Therefore, we draw the empirical conclusion that clustering accuracy and stability can be further improved with CgC by involving the couplings of clusterings. The improvement rate is dependent on the accuracy of base clusterings.

(3) Object-based Comparison

The evaluations (i.e. AC, NMI and CSI) of distinct object-based ensemble methods are exhibited in Fig. 7. Eight data sets with smaller size are chosen because of the high computational complexity

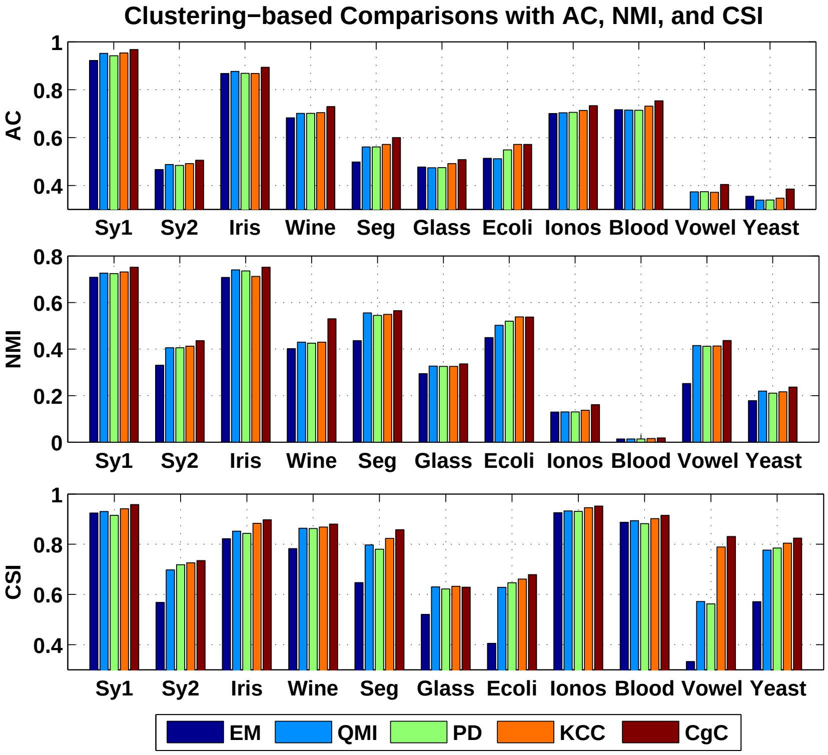


Fig. 6. Clustering-based comparisons.

in this group of experiments. 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 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 VII, 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 the maximum values in the base clustering, e.g. Sy2 and Iris. It is also noteworthy that the evaluation scores of the 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 but from the perspective of information theory, and it lacks the concern of the coupling between objects. However, *ROCK* as a categorical clustering algorithm also leads to poor performance in the clustering ensemble, since it only focuses on the interaction between objects but overlooks the relationship between base clusterings. This discovery is also evidenced by the evaluation results quantified by the AC measure. Statistical testing supports the results on AC and NMI that *OC-Ia* and *OC-H* do not perform worse than *CSPA*, *ROCK*, and *LIMBO*, at a 95% significance level.

Thus, the clustering quality can be enhanced by the involvement of both intra-coupling between objects (e.g. OC-Ia) and inter-coupling between objects (e.g. OC-H) with the latter performing slightly better.

(4) Cluster-based Comparison

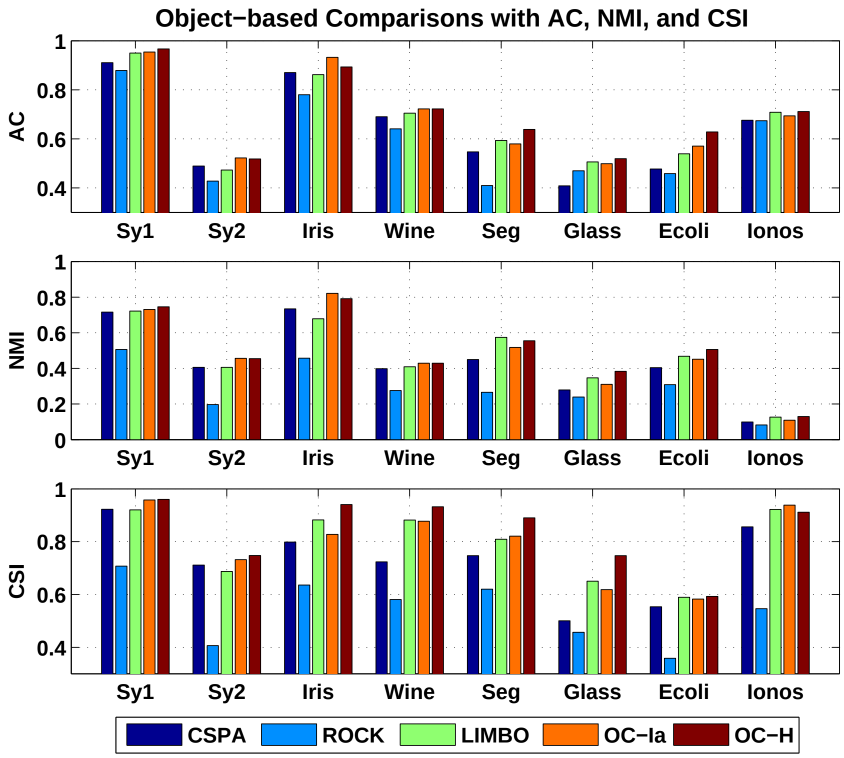


Fig. 7. Object-based comparisons.

Table VIII. 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**

Table VIII reports the experimental results with the cluster-based ensemble methods by using the 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 the table indicates, our proposed *CrC-Ia* and *CrC-C* mostly hold 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%. In addition, the average AC, NMI, and CSI scores of *CrC-Ia* and *CrC-C* across all the data sets are larger than those of comparative approaches and are presented in the last column of Table VIII. Thus, both *CrC-Ia* and *CrC-C* are more robust than other alternatives. 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 being even greater than the relevant maximal scores in base clustering, e.g., Sy2 and Wine. All the results on AC and NMI are supported by a statistical significance test at a 95% significance level.

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 relatively consistent grouping of objects, the chances of encountering a situation where half of the base clusterings put two objects in the same cluster while the other half separates them into 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 VII, respectively. However, for a majority of cases, different base clusterings result in a range of results. Thus, *CrC-C* in particular is expected to demonstrate better performance when differentiating those questionable objects, compared to *CrC-Ia*. We will verify this assumption in detail in Section 6.4.

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

6.4. Data Characteristics and Performance

Building on the previous quality assessments, here we discuss the data characteristics and performance of our proposed *CCE*. Specifically, we address the two assumptions in the previous sections: we aim to discover how the quality of base clusterings affects final clustering accuracy, and how the consistency of base clustering results improves consensus accuracy. Thus, we develop another two groups of experiments to explore the relationship between the data characteristics of base clusterings and the degree of improvement in the final clustering quality.

(1) Quality of Base Clusterings vs Improvement

The first descriptive indicator of data characteristics for base clusterings exhibits the quality of those base clusterings. Here, we use the average AC (i.e. accuracy) or average NMI (i.e. normalized mutual information) of base clusterings generated by *k-means* to represent this indicator to show the quality of base clusterings. In terms of the improvement, the AC performance gain is regarded as the increased proportion of accuracy for *CgC* against the best results among the other three methods (i.e., *EM*, *QMI*, *PD*) considered in Fig. 6, while the NMI performance gain is described as the increased percentage of NMI for *OC-Ia* against the better results between *CSPA* and *ROCK* compared in Fig. 7. Note that these ratios are the relative difference value between two evaluation scores, which is different from the improvement rate in Section 6.3. Formally, the performance gain is defined as:

$$\text{Performance Gain} = [\tau(*) - \tau(Best)]/\tau(Best), \tag{25}$$

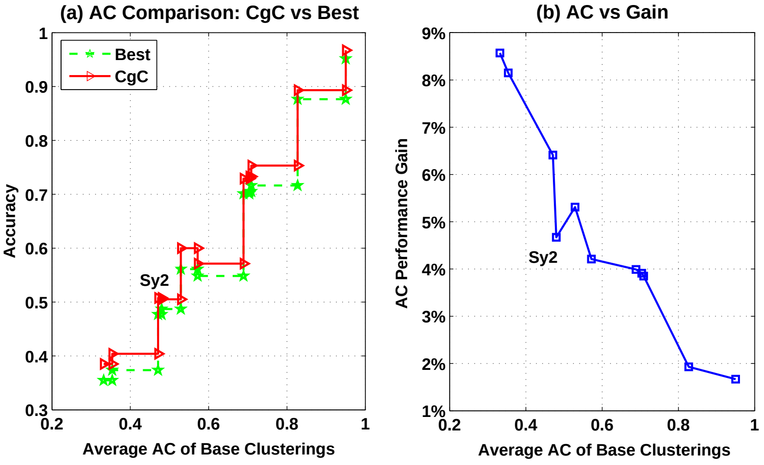


Fig. 8. Quality of base clusterings and AC performance gain for Fig. 6.

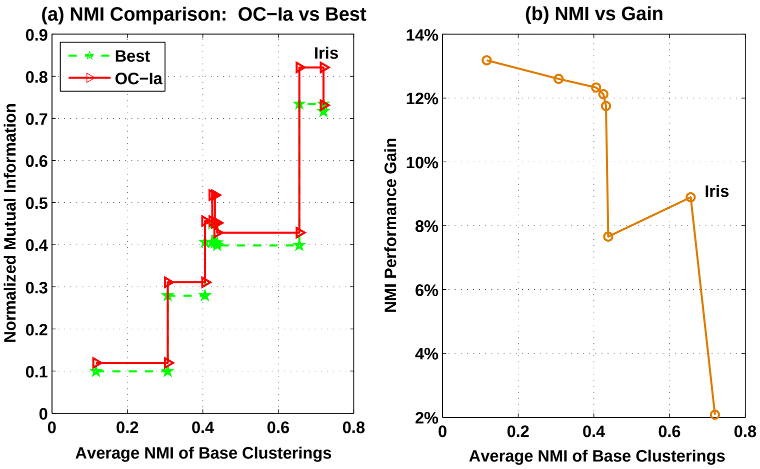


Fig. 9. Quality of base clusterings and NMI performance gain for Fig. 7.

where τ is either AC or NMI as required, $*$ is the proposed method (e.g. *CgC* or *OC-Ia*), and $Best$ is the best comparable algorithm (e.g. $Best \in \{EM, QMI, PD\}$ or $Best \in \{CSPA, ROCK\}$). $\tau(*)$ and $\tau(Best)$ represent the corresponding τ evaluation scores of $*$ and $Best$, respectively.

The results of the relationship between quality and performance gain are reported in Fig. 8 and Fig. 9, which correspond to Fig. 6 and Fig. 7, respectively. Fig. 8(a) shows the staircase chart on AC of *CgC* and the best algorithm among *EM*, *QMI* and *PD*. As can be clearly seen from Fig. 8(b), the larger the average AC of base clusterings (axis x), the smaller the AC performance gain (axis y), for most cases. The only exception is Sy2. This is probably due to the fact that the synthetic data set Sy2 is generated with additional noise, besides which, the Pearson's correlation coefficient between these two variables (i.e. AC of base clusterings and AC performance gain) is -0.9486 with p-value 0.8626×10^{-5} (< 0.05), which means the correlation is negative at a 95% significance level. We draw the same conclusion if we consider NMI values.

Similarly, Fig. 9(a) displays the staircase chart on NMI of *OC-Ia* and the better algorithm between *CSPA* and *ROCK*. Further, Fig. 9(b) reveals that with the exception of Iris, the larger the average NMI of base clusterings (axis x), the smaller the NMI performance gain (axis y). The great variation of NMI for Iris, which is reflected in Table VII with maximal NMI value 0.750 and minimal NMI value 0.427, probably leads to this exception. The corresponding Pearson's correlation coefficient here is -0.7953 between two variables: NMI of base clusterings and NMI performance gain, with p-value 0.0183 (< 0.05). It is also revealed that these two variables are significantly associated in anti-correlation at a 95% significance level. Similar results can be obtained when AC scores are concerned instead.

We have therefore verified our first assumption proposed in Section 6.3. We conclude that the performance gain brought by the coupling of base clusterings against other ensemble methods is negatively associated with the quality of base clusterings, and the result is statistically significant. Intuitively, this conclusion is easy to understand, since the improvement space will automatically become smaller when the base clusterings have already exhibited better quality.

(2) Consistency of Base Clusterings vs Improvement

The consistency of base clusterings is selected as another descriptive indicator of data characteristics for base clusterings. The consistency here describes the variation of clustering results among base clusterings. As pointed out in Section 6.2, CSI reflects the deviation of clustering results across different runs. Thus, we use the CSI of base clusterings (i.e. the last column in Table VII) to represent and quantify the consistency of these results. The larger the CSI, the more consistent the clustering results. Similar to the above Section 6.4, AC and NMI performance gains are again adopted to measure the improvement of *CrC-C* upon *CrC-Ia* in Table VIII. Here, $*$ is *CrC-C* and $Best$ is *CrC-Ia* in Equation (25).

The results obtained for the dependency between consistency and performance gain are presented in Fig. 10. In detail, Fig. 10(a) and Fig. 10(b) exhibit the staircase charts on AC and NMI of *CrC-C* and *CrC-Ia*, respectively. In Fig. 10(c), it is clearly observed that both curves, whether they are AC or NMI, have a general tendency to decrease. That is to say, for most cases, the larger the CSI of the base clusterings (axis x), the smaller the AC or NMI performance gain (axis y). This also means that the performance gain of *CrC-C* upon *CrC-Ia* is associated with the consistency of the base clusterings. If the initial base clusterings have more controversial objects for the final grouping, *CrC-C* is more likely to further refine *CrC-Ia* with the inconsistency. Otherwise, *CrC-C* obtains more or less the same clustering results as *CrC-Ia*; sometimes the results of *CrC-C* are even worse than those of *CrC-Ia*. For instance, there are several points located around the horizontal line of 0% in Fig. 10(c) . Moreover, the Pearson's correlation coefficient between consistency of base clusterings and AC performance gain is -0.9615 with p-value 0 (< 0.05), and the coefficient between consistency and NMI performance gain is -0.8912 with p-value 0.0002 (< 0.05). The statistical test guarantees that the variables of consistency and performance gain are correlated by the negative dependency, significantly with a confidence level at 95%.

The second hypothesis raised in Section 6.3 has consequently been confirmed. Thus, the performance gain caused by the inter-coupling of objects against other ensemble methods is negatively dependent on the consistency of base clustering results, and this consequence is statistically significant. This conclusion explains that if the initial base clusterings have a relatively high level of inconsistency, a further improvement is necessary by also involving the inter-coupling of objects. This conclusion also conforms to the viewpoint proposed by Kuncheva and Hadjitodorov [Kuncheva and Hadjitodorov 2004] as well as Iam-On [Iam-On et al. 2011]: a more accurate partition can be obtained from a diverse ensemble than from the non-diverse case. Here, the diverse ensemble corresponds to the less consistent base clusterings.

In all, we draw the following four conclusions to address the research questions proposed in Section 1: 1) Our proposed similarity measures incorporate the couplings of base clusterings and objects, and have an impressive capacity to discover the implicit relationships in the data. 2) Base clusterings are indeed coupled with each other, and the consideration of such couplings can result in best clustering quality; 3) The inclusion of coupling between objects further improves clustering accuracy and stability; 4) The improvement level or performance gain brought by the coupling of base clusterings is negatively associated with the quality of base clusterings, while the further improvement degree or performance gain caused by the inter-coupling of objects is inversely dependent on

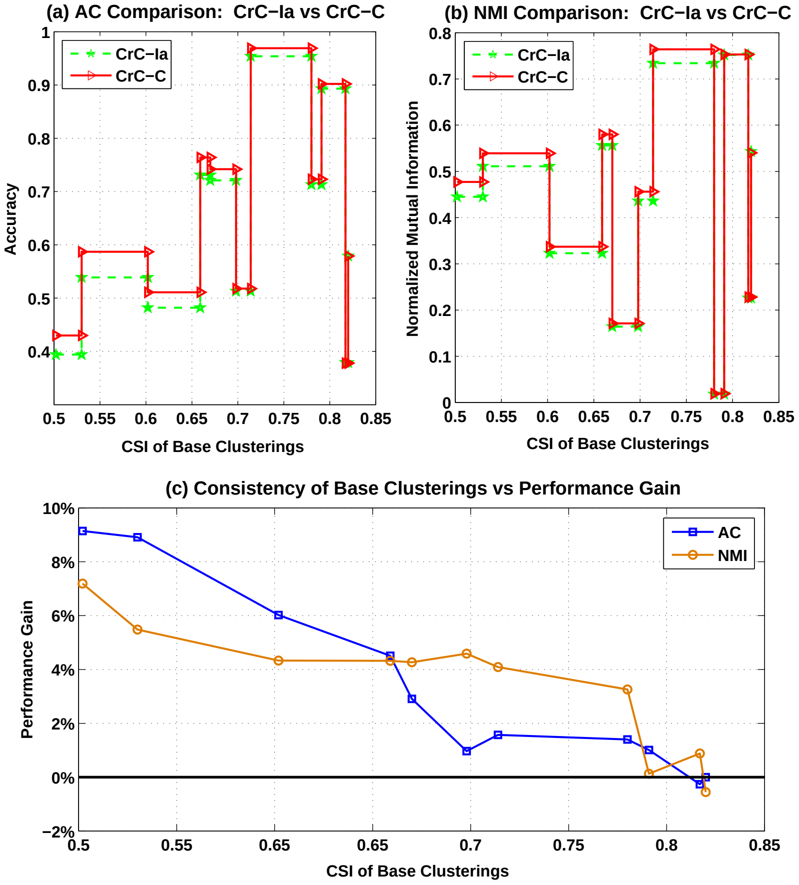


Fig. 10. Consistency of base clusterings and performance gain for Table VIII.

the consistency of the base clustering results. All the results are accordingly supported by statistical tests.

6.5. Applications in Document Clustering

In this part, we present the performance of our proposed *CCE* models in the application of document clustering with large data sets in high dimensions.

6.5.1. Document Clustering Comparison. We conduct experiments on four document collections: D_1 and D_2 are subsets of the 20 newsgroups document collection, D_3 is a subset of the Reuters-21578 document collection, and D_4 is a subset of the WebKB document collection. Detailed information is described as follows:

- D_1 is a subset of 20 newsgroups (20NG) with 1,864 documents and 16,516 terms across five classes. The 20NG document collection [Lang 1995] which consists 20,000 newsgroup documents across 20 classes, is a benchmark data set for document clustering.
- D_2 is the mini-newsgroups version, which has 1,989 documents with 24,809 terms across all the 20 classes in 20NG document collection.
- D_3 is a subset derived from the Reuters-21578 [Lewis 1997], which is a widely used benchmark document collection. This data set has 2,091 documents with 8,674 terms belonging to 8 classes.

Table IX. Document Clustering on AC, NMI and CSI

Data Set		*BOW*	*BOW+KCC*	*BOW+CgC*	*BOW+OC-H*	*BOW+CrC-C*	*CVM*	*CVM+KCC*	*CVM+CgC*	*CVM+OC-H*	*CVM+CrC-C*
AC	D_1	*0.293*	0.402	0.450	**0.463**	0.458	*0.359*	0.512	0.621	**0.639**	**0.639**
	D_2	*0.125*	0.289	**0.362**	0.350	**0.362**	*0.121*	**0.363**	0.360	0.355	**0.363**
	D_3	*0.740*	0.786	0.806	0.836	**0.838**	*0.800*	0.861	0.932	**0.951**	0.936
	D_4	*0.431*	0.586	0.642	**0.681**	0.679	*0.522*	0.659	0.702	0.736	**0.745**
NMI	D_1	*0.139*	0.365	0.362	0.359	**0.375**	*0.182*	0.398	0.398	0.425	**0.436**
	D_2	*0.207*	0.356	**0.411**	0.405	0.409	*0.212*	0.389	**0.456**	0.449	0.445
	D_3	*0.421*	0.615	0.639	**0.662**	0.659	*0.474*	0.679	0.718	**0.742**	**0.742**
	D_4	*0.094*	0.152	0.195	**0.205**	0.201	*0.249*	0.265	0.312	0.359	**0.363**
CSI	D_1	*0.521*	0.786	0.821	0.833	**0.845**	*0.412*	0.801	0.854	**0.859**	0.837
	D_2	*0.436*	0.800	0.803	0.800	**0.812**	*0.501*	**0.829**	0.822	0.828	**0.829**
	D_3	*0.632*	**0.855**	0.853	0.820	0.839	*0.611*	0.810	0.813	0.773	**0.843**
	D_4	*0.325*	0.779	0.792	0.787	**0.826**	*0.402*	0.795	0.801	0.793	**0.825**

- D_4 consists of 4,087 web pages with 7,769 terms classified into four categories, and is a subset of the WebKB document collection which is collected by the WebKB project of the CMU text learning group [Craven et al. 1998].

We use two existing document clustering approaches to compare with the recent consensus clustering model *KCC* [Wu et al. 2015] and our proposed *CCE* framework. The first method is the Bag of Words model (*BOW* for short) [Aizawa 2003] that represents each document as a vector of distinct terms that appear in the document set. Each component of the vector stands for the weight of each term in the document set, and the weight is usually calculated by using the tf-idf weighting scheme. The other one is the context vector model (*CVM* for short) [Billhardt et al. 2002] that applies term co-occurrence pattern to estimate term dependency, and integrates the semantic information with document representation for the calculation of the document similarity. Both approaches are the document representation models, the classic k-means algorithm is employed for document clustering in this group of experiments.

The first set of experiments is done by using *BOW* and k-means to produce the base clusterings with ensemble size $L = 10$. After that, our proposed *CCE* framework including *CgC*, *OC-H*, and *CrC-C*, as well as the *KCC* model [Wu et al. 2015] proposed by Wu et al., is applied on those 10 base clustering results. The second set of experiments is carried out by adopting *CVM* and k-means to generate 10 base clusterings, and then *KCC*, *CgC*, *OC-H*, and *CrC-C* are used on them. We still apply the AC, NMI, and CSI measures to evaluate the quality of document clustering. The larger those scores, the better the model. The experimental results are displayed in Table IX. The highest measure scores of each experimental setting are highlighted in boldface. As can be observed in Table IX, our proposed *CCE* framework can greatly enhance the document clustering quality (i.e. larger AC and NMI) with more stable result (i.e. larger CSI). The performance of *BOW+OC-H* and *BOW+CrC-C* is slightly better than that of *BOW+CgC*. The same performance applies to *CVM*, and *CVM* performs better than *BOW* in most cases. Another observation is that our proposed *CCE* framework generally performs better than *KCC* in the application of document clustering. Statistical testing supports all our findings, at a 95% significance level.

6.5.2. Efficiency Study. We then compare our proposed *CgC*, *OC-H*, *CrC-C* with *KCC* in terms of execution efficiency. Note that only the execution time consumed by clustering ensemble is recorded here. The time used for building base clusterings with *BOW* model or *CVM* model is not counted here, since we focus on the efficiency of clustering ensemble process, rather than that of the base clustering production process.

As indicated in Table X, *KCC* and *CgC* have comparable execution time in clustering ensemble. As we know, our proposed *CgC* considers the coupling relationship between base clusterings, while *KCC* does not address any interdependence. According to Table IX, *CgC* outperforms *KCC* in most cases with respect to the algorithm effectiveness. So, we can claim that *CgC* is better than *KCC* in terms of both effectiveness and efficiency. Another observation from Table X is that *OC-H*

Table X. Comparison of Execution Time (in Seconds)

Model \ Data	D_1	D_2	D_3	D_4
KCC	8.56	9.32	13.98	30.87
CgC	8.65	9.31	13.63	31.12
OC-H	17.54	19.69	30.12	60.59
CrC-C	16.35	19.95	31.20	60.41

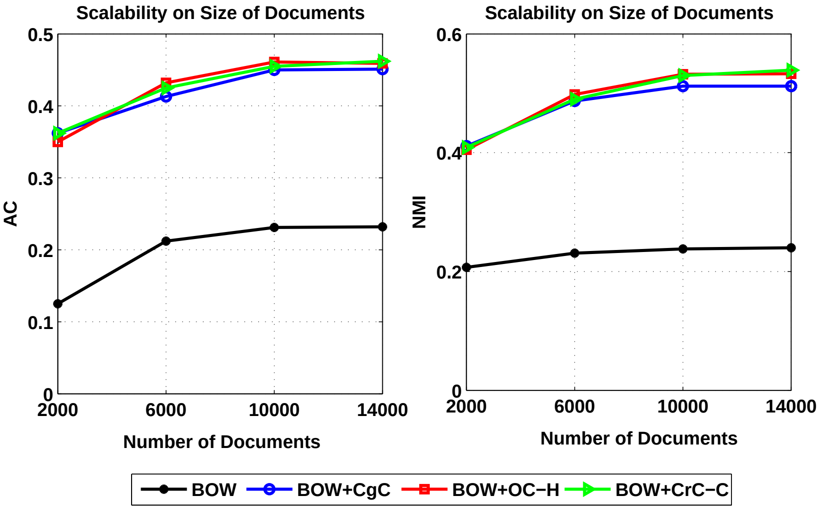


Fig. 11. Scalability Study for Document Clustering

and *CrC-C* have similar computation time, but are a little bit more time-consuming than *KCC* and *CgC*. Apparently, the reason is that *OC-H* and *CrC-C* also take into account the interdependence of objects, which incur more time in modeling such coupling relationship between objects. However, the more time consumed is generally compensated with the higher accuracy induced, based on the results in Table IX. Hence, there is a tradeoff between efficiency and effectiveness for those models. It is very encouraging that the execution time is still acceptable, not too large, even for *OC-H* and *CrC-C*.

6.5.3. Scalability Study. To further evaluate the performance of our proposed models on large document collection, we conduct a set of experiments on the 20 newsgroups document collection by increasing the number of documents from 2,000 to 14,000 with an increment of 4,000 new documents.

Fig. 11 shows how AC and NMI scores vary with the size of document collection. Both scores indicate that the clustering quality is not affected by the size of document collection. The improvement over *BOW* is close to that in Table IX. Thus, we can say that our proposed models work well on large document collection. The results demonstrate that our proposed framework is quite a practical approach. Besides, our model can be easily and effectively parallelized, leading to faster execution time on large document collection when given an available parallel system.

6.6. Comparisons on Data with Large Size and Dimensionality

In this part, we perform experiments on data sets with large size and dimensionality to show the effectiveness of our proposed framework. The basic statistical characteristics of such data sets are described in Table XI, with the data size ranging from 2,521 to 581,012, and the data dimensionality ranging from 9 to 126,373. Four of the data sets are from UCI repository [Frank and Asuncion 2010],

Table XI. Data Sets with Large Size and Dimensionality

Data Set	m	n	t^p	Source
Satimage	4,435	36	6	UCI
Pendigits	10,992	16	10	UCI
Shuttle	58,000	9	7	UCI
Covtpe	581,012	54	7	UCI
MM	2,521	126,373	2	TREC
Reviews	4,069	126,373	5	TREC
La12	6,279	31,472	6	TREC
Sports	8,580	126,373	7	TREC

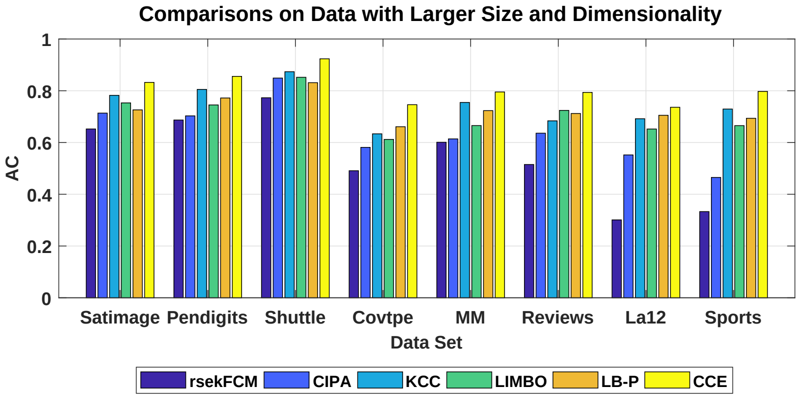


Fig. 12. Comparisons on Data with Larger Size and Dimensionality on AC

and another four are from TREC repository [Wu et al. 2015]. In Table XI, m represents the number of objects, n stands for the number of attributes, and t^p is the number of predefined labels.

We compare our proposed coupled framework *CCE* with three clustering ensemble algorithms and two recent clustering models. Those three clustering ensemble methods are selected from Fig. 6, Fig. 7, and Table VIII. In other words, *KCC*, *LIMBO*, *LB-P* are the representatives of clustering-based, object-based, and cluster-based approaches. We also compare with two recent clustering algorithms, they are the prototype-based kernel algorithm *rsekFCM* [Havens et al. 2012] and the clustering by iterative partitioning and point attractor representation model *CIPA* [Shao et al. 2016]. For our proposed *CCE*, we apply the "big data issue" solution proposed in Section 5.3 to speed up the whole process. For the state-of-the-art models, we implement with their default settings.

The results are shown in Fig. 12 by using accuracy (AC) to measure the effectiveness of clustering. For our proposed *CCE*, we report the best result among *CgC*, *OC-Ia*, *OC-H*, *CrC-Ia*, and *CrC-C* in terms of accuracy. As demonstrated in Fig. 12, *CCE* performs the best across all the eight data sets, with KCC being the second, and then with *LB-P* being slightly better than *LIMBO*. Another interesting observation is that all the clustering ensemble algorithms outperform the clustering algorithms. This is because the clustering ensemble is conducted on the initial clusterings to improve their results by working on multiple clustering runs. We can also see that *CIPA* performs better than *rsekFCM* on all the data sets in terms of accuracy, which is consistent with the conclusion in [Shao et al. 2016]. Similar results apply for the NMI comparisons. Statistical testing supports all our results, at a 95% significance level.

7. DISCUSSIONS

Below, we discuss the potential and future opportunities related to our proposed *CCE* from two aspects. The depth aspect discusses the extension of current definitions on coupling, while the width aspect explores other approaches apart from the consensus function based clustering ensemble and other stages in the process of the clustering ensemble.

Depth Aspect: According to the conclusion in Section 6.4, the improvement on clustering performance is largely dependent on the data characteristics of base clusterings, which are quantified as the quality and the consistency of base clusterings. Hence, we need to consider these two descriptive indicators in the coupled clustering ensemble.

In our implementation, we regard the weight λ_k of base clustering bc_k in Equation (4) to be the same, i.e., $\lambda_k = 1/(L-1)$, where L is the number of base clusterings. However, the clustering quality (e.g., AC and NMI) of each base clustering, denoted as q_k, can be adapted to substitute λ_k to differentiate the contributions made by distinct base clusterings. Here, we normalize q_k by $q'_k = q_k / \sum_{k=1,k\neq j}^{L} q_k$ to make $\sum_{k=1,k\neq j}^{L} q'_k = 1$ to satisfy the requirement for λ_k. In this way, base clustering that performs better will contribute more in the calculation of similarity between cluster labels. Therefore, we can incorporate the indicator q'_k on the quality of base clusterings into our method.

In Equation (10), the ratio of the number of common θ-neighbors is used to measure the similarity between two objects. However, the consistency $\mu \in [0, 1]$ (e.g., CIS) of all the base clusterings can be utilized to control the extent to which include the inter-coupling of objects. The purpose here is to adjust the effect of inter-coupled objects according to the consistency of base clusterings. We can then alternatively replace $\delta^{CO}(u_x, u_y|U)$ with $\mu \cdot \delta^{IaO}(u_x, u_y) + (1-\mu) \cdot \delta^{CO}(u_x, u_y|U)$. Thus, the inter-coupling of objects will be less emphasized when the base clusterings obtain more approximate results. If all the base clusterings ideally perform the same (i.e., $\mu = 1$), the similarity between objects degenerates to δ^{IaO}. Consequently, we can involve the indicator μ on the consistency of base clusterings into our model.

These two descriptive indicators adapt our *CCE* to a soft version *S-CCE*, since *S-CCE* considers how much contribution each base clustering makes and to what extent we involve the inter-coupling of objects. However, the previous indicator requires the label information at the stage of generating base clusterings. If this information is unavailable during the whole process of the clustering ensemble, then only the second indicator can be used.

Width Aspect: As mentioned in Section 2, there are three ways to aggregate the base clusterings: consensus functions, categorical clusterings, and direct optimizations. We mainly focus on the consensus function based clustering ensemble to propose the coupled clustering *CCE*. For the second option on categorical clusterings, we have designed the coupled nominal similarity in unsupervised learning [Wang et al. 2015], which induces alternatives to cluster categorical data and also forms a part of *CCE*, besides which, we have already involved the widely used categorical clustering algorithms (i.e., *ROCK* [Guha et al. 2000] and *LIMBO* [Andritsos et al. 2004]) in our experiments. The third group of methods on direct optimizations selects candidates among all the clusters produced by base clusterings and then adjusts them to achieve the minimal cost [Christou 2011]. They totally ignore the consensus of base clusterings, and do not rely on the similarity or distance between base clusterings, objects, and clusters. Thus, our proposed *CCE* does not fit the direct optimizations based clustering ensemble. In addition, the direct optimizations based approaches require the detailed information of each object (i.e. the attribute values) to obtain the sum of intra-cluster distance as the cost of each cluster, while *CCE* still works well despite the lack of such prior information and enables the privacy-preserving and distributed mode of data analysis.

Also as introduced in Section 2, the whole process of the clustering ensemble is composed of three stages: building base clustering, aggregating base clusterings, and post-processing clustering. In this paper, our proposed *CCE* is constructed for the second stage, and the first and last stages are fixed as in comparative methods. In reality, base clusterings and post-processing techniques are also shown to affect the performance of clustering ensemble [Iam-On et al. 2011]. In our experiments, *k-means* on random sub-sampling with a fixed k is adopted to build base clusterings, and homogeneous results are accordingly obtained. Alternatively, different values of k can be selected, and distinct approaches are also expected to generate heterogeneous base clusterings. The input base clusterings then exhibit a higher level of diversity than those we have used. Note that the consistency pointed out in Section 6.4 is just one aspect of diversity among base clusterings. At the post-processing stage, three fundamental clustering algorithms are employed, namely, *k-means*, *agglomerative algorithm*, and *METIS*. However, advanced similarity or distance based clustering algorithms, such as spectral clustering [Luxburg 2007] and affinity propagation [Frey and Dueck 2007], can be applied to further improve the quality of the clustering ensemble. In future studies, therefore, we will also examine the heterogeneous structure of base clusterings and advanced post-processing clustering algorithms in our proposed *CCE* to enhance the performance of the whole process.

8. CONCLUSION

The clustering ensemble has been introduced as a more accurate alternative to individual (base) clustering algorithms. Existing approaches are mostly based on the IIDness assumption, which is too restrictive to explore their maximum potentials. This paper has proposed the coupled clustering ensemble, i.e. *CCE*, to incorporate various interactions between base clusterings and between objects, which constitute the behavior interior dimensions. The key contributions are as follows:

- The interdependent nature is described from the perspectives of clustering-based, object- based, and cluster-based algorithms, and reveals that they are essential to the clustering ensemble.
- Both the couplings between base clusterings and between data objects are addressed in *CCE* to support the integrated interdependence relationships.
- We propose several similarity measures that incorporate the couplings of base clusterings and objects, and they exhibit an impressive ability to capture the implicit relationships in data.
- Our proposed *CCE* is evaluated against nine existing clustering ensemble methods and four clustering algorithms on various benchmark data sets in terms of accuracy, stability, robustness, and statistical significance.
- We empirically explore the relationship between the data characteristics of base clusterings and the degree of possible improvement in the final clustering quality.
- The applications in document clustering, as well as on the data sets with large size and dimensionality, further demonstrate the effectiveness, efficiency, and scalability of our *CCE* models.

In the future, we may consider the following issues to further expand our current work. How should we fix the weights λ_k of base clustering bc_k in *IeCSC* rather than simply treating them equally? Further, should we introduce a weight to control the couplings of objects during the process of the clustering ensemble? Is there any other way to model the coupling of objects by considering the relative common neighborhood rather than the absolute θ-neighborhood? How do we fix the number of final clusters? We will work on these issues, as mentioned in the discussions, and will also analyze the heterogeneous structure of base clusterings and the advanced post-processing clustering techniques in *CCE*. Finally, we will consider the coupling of clusters and then extend this coupled idea to the supervised learning process.

9. ACKNOWLEDGEMENT

The authors would like to thank the associate editor and all the anonymous reviewers for their invaluable comments on our paper.

REFERENCES

Akiko Aizawa. 2003. An Information-Theoretic Perspective of TF-IDF Measures. *Information Processing & Management* 39, 1 (2003), 45–65.

P. Andritsos, P. Tsaparas, R.J. Miller, and K.C. Sevcik. 2004. LIMBO: Scalable clustering of categorical data. In *Proceedings of the 9th International Conference on Extending Database Technology*. 123–146.

Holger Billhardt, Daniel Borrajo, and Victor Maojo. 2002. A Context Vector Model for Information Retrieval. *Journal of the American Society for Information Science and Technology* 53, 3 (2002), 236–249.

S. Boriah, V. Chandola, and V. Kumar. 2008. Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the 8th SIAM International Conference on Data Mining*. 243–254.

D. Cai, X. He, and J. Han. 2005. Document clustering using locality preserving indexing. *IEEE Transactions on Knowledge and Data Engineering* 17, 12 (2005), 1624–1637.

Longbing Cao. 2014. Non-iidness learning in behavioral and social data. *Comput. J.* 57, 9 (2014), 1358–1370.

Longbing Cao, Yuming Ou, and Philip S Yu. 2012. Coupled behavior analysis with applications. *IEEE Transactions on Knowledge and Data Engineering* 24, 8 (2012), 1378–1392.

I.T. Christou. 2011. Coordination of cluster ensembles via exact methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 2 (2011), 279–293.

S. Cost and S. Salzberg. 1993. A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning* 10, 1 (1993), 57–78.

Mark Craven, Andrew McCallum, Dan PiPasquo, Tom Mitchell, and Dayne Freitag. 1998. *Learning to Extract Symbolic Knowledge from the World Wide Web*. Technical Report.

C. Domeniconi and M. Al-Razgan. 2009. Weighted cluster ensembles: methods and analysis. *ACM Transactions on Knowledge Discovery from Data* 2, 4 (2009), 17.

Xiaoli Zhang Fern and Carla E. Brodley. 2004. Solving cluster ensemble problems by bipartite graph partitioning. In *Proceedings of the 21st International Conference on Machine Learning*. 36–43.

Lucas Franek and Xiaoyi Jiang. 2014. Ensemble clustering by means of clustering embedding in vector spaces. *Pattern Recognition* 47, 2 (2014), 833–842.

A. Frank and A. Asuncion. 2010. *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences.

B.J. Frey and D. Dueck. 2007. Clustering by passing messages between data points. *Science* 315, 5814 (2007), 972–976.

G. Gan, C. Ma, and J. Wu. 2007. *Data Clustering: Theory, Algorithms, and Applications*. ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA.

Guojun Gan and Michael Kwok-Po Ng. 2015. Subspace clustering using affinity propagation. *Pattern Recognition* 48, 4 (2015), 1455–1464.

César García-Osorio, Aida de Haro-García, and Nicolás García-Pedrajas. 2010. Democratic instance selection: a linear complexity instance selection algorithm based on classifier ensemble concepts. *Artificial Intelligence* 174, 5 (2010), 410–441.

A. Gionis, H. Mannila, and P. Tsaparas. 2007. Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data* 1, 1 (2007), 1–30.

S. Guha, R. Rastogi, and K. Shim. 2000. ROCK: A robust clustering algorithm for categorical attributes. *Information Systems* 25, 5 (2000), 345–366.

Timothy C Havens, James C Bezdek, Christopher Leckie, Lawrence O Hall, and Marimuthu Palaniswami. 2012. Fuzzy c-means algorithms for very large data. *IEEE Transactions on Fuzzy Systems* 20, 6 (2012), 1130–1146.

Natthakan Iam-On and Tossapon Boongoen. 2012. Improved link-based cluster ensembles. In *Proceedings of the International Joint Conference on Neural Networks 2012*. 1–8.

N. Iam-On, T. Boongoen, S. Garrett, and C. Price. 2011. A link-based approach to the cluster ensemble problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 12 (2011), 2396–2409.

Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. 2009. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data* 3, 1 (2009), 1–58.

L.I. Kuncheva and S.T. Hadjitodorov. 2004. Using diversity in cluster ensembles. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, Vol. 2. 1214–1219.

L.I. Kuncheva and D.P. Vetrov. 2006. Evaluation of stability of k-means cluster ensembles with respect to random initialization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 11 (2006), 1798–1808.

Ken Lang. 1995. Newsweeder: Learning to Filter Netnews. In *Proceedings of the 12th International Conference on Machine Learning*. 331–339.

David D Lewis. 1997. Reuters-21578 Text Categorization Test Collection, Distribution 1.0. *http://www. research. att. com/l̃ewis/reuters21578. html* (1997).

T. Li, M. Ogihara, and S. Ma. 2010. On combining multiple clusterings: an overview and a new perspective. *Applied Intelligence* 33, 2 (2010), 207–219.

D. Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*. 296–304.

Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, Junjie Wu, and Sen Wu. 2013. Understanding and Enhancement of Internal Clustering Validation Measures. *IEEE Transactions on Cybernetics* 43, 3 (2013), 982–994.

U. Luxburg. 2007. A tutorial on spectral clustering. *Statistics and Computing* 17, 4 (2007), 1–32.

N. Nguyen and R. Caruana. 2007. Consensus clusterings. In *Proceedings of the 7th IEEE International Conference on Data Mining*. 607–612.

K. Punera and J. Ghosh. 2007. Soft cluster ensembles. In *Advances in Fuzzy Clustering and Its Applications*, Jose Valente de Oliveira and Witold Pedrycz (Eds.). John Wiley & Sons Ltd, Chichester, UK, 69–91.

Junming Shao, Qinli Yang, Hoang-Vu Dang, Bertil Schmidt, and Stefan Kramer. 2016. Scalable clustering by iterative partitioning and point attractor representation. *ACM Transactions on Knowledge Discovery from Data* 11, 1 (2016), 5:1–5:23.

A. Strehl and J. Ghosh. 2002. Cluster ensembles–a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research* 3 (2002), 583–617.

Zhongbin Sun, Qinbao Song, Xiaoyan Zhu, Heli Sun, Baowen Xu, and Yuming Zhou. 2015. A novel ensemble method for classifying imbalanced data. *Pattern Recognition* 48, 5 (2015), 1623–1637.

A. Topchy, A.K. Jain, and W. Punch. 2005. Clustering ensembles: models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 12 (2005), 1866–1881.

Sandro Vega-Pons and Jose Ruiz-Shulcloper. 2011. A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence* 25, 3 (2011), 337.

C. Wang, L. Cao, M. Wang, J. Li, W. Wei, and Y. Ou. 2011. Coupled nominal similarity in unsupervised learning. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management*. 973–978.

Can Wang, Chi-Hung Chi, Wei Zhou, and Raymond K Wong. 2015. Coupled Interdependent Attribute Analysis on Mixed Data.. In *AAAI*. 1861–1867.

Can Wang, Xiangjun Dong, Fei Zhou, Longbing Cao, and Chi-Hung Chi. 2015. Coupled Attribute Similarity Learning on Categorical Data. *IEEE Transactions on Neural Networks and Learning Systems* 26, 4 (2015), 781–797.

Can Wang, Zhong She, and Longbing Cao. 2013. Coupled clustering ensemble: Incorporating coupling relationships both between base clusterings and objects. In *The 29th International Conference on Data Engineering*. 374–385.

P. Wang, C. Domeniconi, and K. Laskey. 2010. Nonparametric Bayesian clustering ensembles. In *Machine Learning and Knowledge Discovery in Databases*, Hendrik Blockeel Kristian Kersting and Siegfried Nijssen Filip Železnỳ (Eds.). Springer, Heidelberg, Germany, 435–450.

D.H. Wolpert and W.G. Macready. SFI-TR-95-02-010, 1996. *No free lunch theorems for search*. Technical Report. Citeseer.

Junjie Wu, Hongfu Liu, Hui Xiong, and Jie Cao. 2013. A Theoretic Framework of K-Means-Based Consensus Clustering. In *The 23rd International Joint Conference on Artificial Intelligence*. 1799–1805.

Junjie Wu, Hongfu Liu, Hui Xiong, Jie Cao, and Jian Chen. 2015. K-Means-Based Consensus Clustering: A Unified View. *IEEE Transactions on Knowledge and Data Engineering* 27, 1 (2015), 155–169.