A Link-Based Approach to the Cluster Ensemble Problem

Natthakan Iam-On, Tossapon Boongoen, Simon Garrett, and Chris Price

Abstract—Cluster ensembles have recently emerged as a powerful alternative to standard cluster analysis, aggregating several input data clusterings to generate a single output clustering, with improved robustness and stability. From the early work, these techniques held great promise; however, most of them generate the final solution based on incomplete information of a cluster ensemble. The underlying ensemble-information matrix reflects only cluster-data point relations, while those among clusters are generally overlooked. This paper presents a new link-based approach to improve the conventional matrix. It achieves this using the similarity between clusters that are estimated from a link network model of the ensemble. In particular, three new link-based algorithms are proposed for the underlying similarity assessment. The final clustering result is generated from the refined matrix using two different consensus functions of feature-based and graph-based partitioning. This approach is the first to address and explicitly employ the relationship between input partitions, which has not been emphasized by recent studies of matrix refinement. The effectiveness of the link-based approach is empirically demonstrated over 10 data sets (synthetic and real) and three benchmark evaluation measures. The results suggest the new approach is able to efficiently extract information embedded in the input clusterings, and regularly illustrate higher clustering quality in comparison to several state-of-the-art techniques.

Index Terms—Clustering, cluster ensembles, cluster relations, link-based similarity, data mining.

1 Introduction

Data clustering is one of the fundamental tools used for understanding the structure of a data set. It has been successfully applied to a variety of problem domains such as biology, customer relationship management, information retrieval, pattern recognition, psychology, and recommender systems. In addition, the recent development of clustering cancer gene expression data has attracted a lot of interest among computer scientists, biological, and clinical researchers during the past decade. Clustering aims to categorize data objects into groups or clusters such that the objects in the same cluster are more similar to each other than to those in different clusters. Although a large number of clustering algorithms have been introduced in the literature [24], there is no single clustering algorithm that performs best for all data sets [30], i.e., unable to discover all types of cluster shapes and structures presented in data [9], [17], [49]. Each algorithm has its own strengths and weaknesses. For a particular set of data, different algorithms, or even the same algorithm with different parameters, usually provide distinct solutions. Therefore, it is extremely difficult for users to decide which algorithm would be the "proper" alternative.

Recently, cluster ensembles have emerged as an effective solution that is able to overcome these limitations and improve the robustness as well as the quality of clustering results. Several papers have been published here [17], [31], [48] that have helped to develop this field. The main objective of cluster ensembles is to combine different decisions of various clustering algorithms in such a way as to achieve accuracy superior to those of individual clustering. Examples of well-known ensemble methods are:

1. the feature-based approach that treats the problem of cluster ensembles as the clustering of categorical data (i.e., cluster labels) [5], [6], [40], [47], [48],
2. the direct approach that finds the final partition through relabeling the base clustering results [13], [19],
3. the graph-based approach that employs the graph representation and partitioning technique [7], [12], [44], and
4. the pairwise similarity approach that makes use of co-occurrence relationships between all pairs of data points [3], [11], [15], [16], [17], [38].

Despite their theoretical and practical contributions, almost all cluster ensemble methods found in the literature make use of information available in an ensemble only at a coarse level. They commonly generate the final result from a knowledge pool (or a metalevel information matrix) which is simply created by stacking up ensemble members' decisions. The relations between these decisions (or data partitions) have been unfortunately overlooked [23]. Very few attempts (e.g., [12] and [44]) have been made to bring in

- *N. Iam-On is with the School of Information Technology, Mae Fah Luang University, Muang, Chiang Rai, 57100, Thailand. E-mail: nt.iamon@gmail.com.*
- *T. Boongoen is with the Royal Thai Air Force Academy, 171/1 Klongthanhon, Saimai, Bangkok 10220, Thailand. E-mail: turtletoss@hotmail.com.*
- *S. Garrett is with Aispire Consulting Ltd., Tanyrallt, Aberystwyth, SY23 3PG, UK. E-mail: s.garrett@aispire.co.uk.*
- *C. Price are with the Department of Computer Science, Aberystwyth University, Llandinam Building, Aberystwyth, Ceredigion, SY23 3DB, UK. E-mail: cjp@aber.ac.uk.*

Manuscript received 9 Oct. 2009; revised 22 Sept. 2010; accepted 3 Mar. 2011; published online 27 Apr. 2011.
Recommended for acceptance by M. Meila.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2009-10-0674.
Digital Object Identifier no. 10.1109/TPAMI.2011.84.

this additional information, but only implicitly via a graph representation scheme. In particular, the relation between graph elements corresponding to clusters in an ensemble is restricted only to a shared-content basis. As such, many associations, especially those among clusters of each data partition, are not addressed and used to their true potential. This shortcoming is reflected by the sparseness of the ensemble-information matrix, which presents only cluster-data point relations, while ignoring those between clusters.

Based on these insightful observations, this paper introduces a new link-based approach to the cluster ensemble problem. It aims to refine the ensemble-information matrix using the similarity between clusters in the ensemble under examination. In particular, three new link-based algorithms are proposed for similarity evaluation from the network model representing the ensemble. Having achieved the refined matrix, two new consensus methods are proposed to derive the ultimate clustering result: 1) feature-based partitioning (FBP) and 2) bipartite graph partitioning (BGP), respectively. Unlike the existing feature-based approach [46], [47] that makes use of a categorical data clustering technique, the FBP method applies numerical clustering algorithms of k-means [21] and PAM [28] to obtain the final data partition. In addition, the BGP extends the graph-based technique of [12]. It transforms the refined matrix into a weighted bipartite graph to which the spectral graph-partitioning algorithm [39] is applied.

The new framework uniquely applies the methodology of link analysis [4], [18], [33] to cluster ensembles. Also, it is largely different from the recent attempts [7], [41] to refine the ensemble-information matrix. To obtain a more fine-grain result for each ensemble member, the weighted distance measure is used to represent a soft relation between a pair of data point and cluster [7]. For the same reason, the method of [41] employs a fuzzy clustering algorithm to create the ensemble. Despite these intuitive ideas, the refined solutions of different base clusterings are stacked up to form the ensemble-information matrix, without addressing the relations among input clusterings.

The rest of this paper is organized as follows: Section 2 presents the cluster ensemble problem upon which the current research has been established. The proposed link-based approach, including the underlying intuition of refining the ensemble-information matrix and details of link-based similarity measures, is described in Section 3. Following that, Section 4 exhibits the empirical evaluation of this new approach against other cluster ensemble algorithms, over real and synthesized data sets. The paper is concluded in Section 5 with suggestions for further work.

2 The Cluster Ensemble Problem

Let $X = \{x_1, \ldots, x_N\}$ be a set of N data points and let $\Pi = \{\pi_1, \ldots, \pi_M\}$ be a cluster ensemble with M base clusterings, each of which is also referred to as an "ensemble member." Each base clustering returns a set of clusters $\pi_g = \{C_1^g, C_2^g, \ldots, C_{k_g}^g\}$ such that $\bigcup_{j=1}^{k_g} C_j^g = X$, where k_g is the number of clusters in the gth clustering. For each $x_i \in X$, $C(x_i)$ denotes the cluster label to which the data point x_i belongs. In particular to the gth clustering, $C(x_i) =$ "j" or "C_j^g" if $x_i \in C_j^g$.

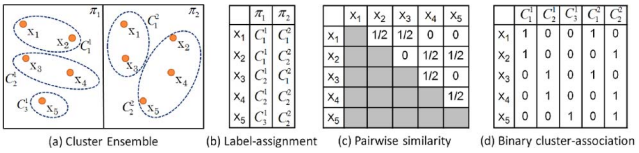


Fig. 1. Examples of (a) cluster ensemble and the corresponding ensemble-information matrices: (b) label-assignment, (c) pairwise similarity, and (d) binary cluster-association, respectively. Note that $X = \{x_1, \ldots, x_5\}$, $\Pi = \{\pi_1, \pi_2\}$, $\pi_1 = \{C_1^1, C_1^1, C_3^1\}$, and $\pi_2 = \{C_1^2, C_2^2\}$.

The problem is to find a new partition π^* of a data set X that summarizes the information from the cluster ensemble Π. This metalevel method involves two major tasks of: 1) generating a cluster ensemble, and 2) producing the final partition (normally referred to as a "consensus function").

It has been illustrated that ensembles are most effective when constructed from a set of predictors whose errors are dissimilar [29]. To a great extent, diversity among ensemble members is introduced to enhance the result of an ensemble [31]. Specific to data clustering, the results obtained with any single algorithm over many iterations are usually very similar. In such circumstance where all ensemble members agree on how a data set should be partitioned, aggregating the base clustering results will show no improvement over any of the constituent members. Several heuristics have been proposed to introduce artificial instabilities in clustering algorithms, hence the diversity within a cluster ensemble. The following ensemble generation methods yield different clusterings of the same data, by exploiting different cluster models and different data partitions.

- *Homogeneous ensembles*. Base clusterings are created using repeated runs of a single clustering algorithm, with several sets of parameter initializations, such as cluster centers of the k-means clustering method [16], [17], [19], [47].
- *Different-k*. One of the most successful technique is randomly selecting the number of clusters (k) for each ensemble member [17], [31].
- *Data subspace/subsample*. A cluster ensemble can also be achieved by applying manifold subsets of initial data to base clusterings. It is intuitively assumed that each clustering algorithm can provide different levels of performance for different partitions of a data set [7]. Practically, data partitions are obtained by projecting data onto different subspaces [11], [46], choosing different subsets of features [44], [50], or data sampling [10], [13], [36].
- *Heterogeneous ensembles*. A number of different clustering algorithms are exploited as base clusterings [3], [22], [32].
- *Mixed heuristics*. In addition to using one of the aforementioned methods, any combination of them can be applied as well [23], [38], [40], [44].

Having obtained the cluster ensemble, a variety of consensus functions have been developed and made available for deriving the final data partition. Each consensus function utilizes a specific form of ensemble-information matrix that summarizes the base clustering results. Given the ensemble of Fig. 1a, three general types of such a matrix are illustrated in Figs. 1b to 1d. The pairwise

similarity matrix shown in Fig. 1c contains the similarity among data points, which can be constructed from the original label-assignment matrix whose example is given in Fig. 1b. Furthermore, as presented in Fig. 1d, the binary cluster-association (BA) matrix provides a cluster-specific view of the original label-assignment matrix. The association degree of a data point belonging to a specific cluster is either 1 or 0. In light of this background, consensus methods can be categorized as follows:

- *Feature-based approach*. It transforms the problem of cluster ensembles to the clustering of categorical data. Each base clustering provides a cluster label as a new feature describing each data point (see Fig. 1b), which is utilized to formulate the final solution [5], [6], [40], [47], [48]. For instance, the technique of [5] makes use of Linear Programming to find a correspondence between the labels of base clusterings and those of the optimal final-clustering. The clustering method for categorical data of [6] employs the genetic algorithm to search for the "median" partition, which is the most similar to the data partitions obtained from base clusterings. In addition, the aggregation of multiple clustering results has been considered as a maximum likelihood estimation problem, and EM algorithms [40], [47], [48] have been proposed for finding the consensus clustering.
- *Direct approach*. This is based on relabeling π_g and searching for the π^* that has the best match with all π_g, $g = 1 \dots M$ [13], [19]. The underlying relabel process allows the homogeneous labels to be established from heterogeneous clustering decisions, where each base clustering possesses a unique set of decision labels (see Fig. 1b).
- *Pairwise similarity approach*. It creates a matrix, containing the pairwise similarity among data points (see Fig. 1c), to which any similarity-based clustering algorithm (e.g., hierarchical clustering) can be applied [3], [11], [16], [17], [38].
- *Graph-based approach*. A number of methods following this approach make use of the graph representation to solve the cluster ensemble problem [7], [12], [44]. Specific to the consensus methods of [7], [44], a graph representing the similarity among data points is created from a pairwise matrix similar to that given in Fig. 1c. To achieve the final clustering result, this graph is divided into a definite number of approximately equal-sized partitions, using METIS [27]. In addition, the binary cluster-association matrix shown in Fig. 1d has also been used for the generation of a bipartite graph whose vertices represent both data points and clusters. According to [12], the solution to a cluster ensemble problem is to divide this graph using either METIS or Spectral graph partitioning (SPEC) [39].

3 A Novel Link-Based Approach

This section introduces a new link-based approach to the cluster ensemble problem. Fig. 2 summarizes the entire process of the proposed framework, which includes two steps of: 1) creating the refined ensemble-information matrix using a link-based similarity algorithm, and 2) generating the final data partition by exploiting two different consensus methods of Feature-Based and Graph-Based Partitioning.

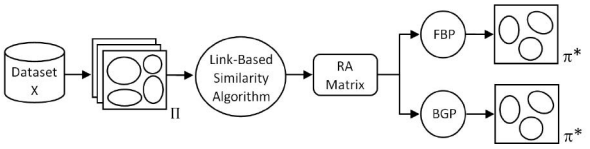


Fig. 2. The framework of the link-based cluster ensemble approach. A link-based similarity algorithm is first applied to create a refined cluster-association (RA) matrix. Then, a final clustering result, π^*, is produced using two consensus functions of Featured-Based Partitioning and Bipartire Graph Partitioning.

3.1 Refining BA Matrix through Cluster Relations

The proposed approach follows several advanced cluster ensemble methods such as [12] and [44], which apply different consensus functions to the binary cluster-association matrix. This metalevel matrix, i.e., $BA \in \{0,1\}^{N \times P}$, where N and P denote the number of data points and clusters in an ensemble, summarizes the cluster-data point relations occurring in the examined ensemble Π. Each entry $BA(x_i, cl) \in \{0,1\}$ represents a "crisp" association degree that a data point $x_i \in X$ has with a cluster $cl \in \Pi$. As presented in Fig. 1d, $BA(x_1, C_1^1) = 1$ and $BA(x_1, C_2^1) = 0$ since the data point x_1 has been labeled as a member of the cluster C_1^1, but not that of the cluster C_2^1. Specific to an ensemble of hard clusterings that is the focus of this research, a data point $x_i \in X$ is a member of only one cluster in any clustering $\pi_g \in \Pi$.

It is shown in the example that the BA matrix is generally sparse, with a large number of entries being "0". Intuitively, this particular characteristic that is commonly encountered with the ensemble of hard clustering results may limit the quality of a data partition generated by any consensus function. In order to resolve this problem, a few methods have been introduced in the literature to obtain a refined information matrix. The approach of [41], namely, "soft cluster ensembles," uses a fuzzy clustering algorithm for the generation of base clusterings. For each clustering $\pi_g \in \Pi$, a data point $x_i \in X$ can belong to several clusters of π_g, each with an association degree between $[0,1]$. As a result, the BA-like matrix can be formed as $BA' \in [0,1]^{N \times P}$. A similar refinement method has also been developed by [7] as part of "weighted cluster ensembles." Unlike the former approach, this method does not utilize a fuzzy clustering technique, but employs hard clustering solutions of the subspace clustering algorithm, namely, "Locally Adaptive Clustering (LAC)" [8]. Specific to each clustering $\pi_g \in \Pi$, the association degree that $x_i \in X$ has with the cluster $cl \in \pi_g$ is the weighted distance measure between x_i and the center of cl, which is normalized by the summation of such measure between x_i and the centers of all clusters in π_g.

Despite the intuition behind these methods, the underlying problem is tackled mainly by refining the results obtained from each base clustering, using a fuzzy algorithm or a weighted distance metric. Once such fine-grain solutions are achieved they are, as with the conventional cluster ensemble approach, simply stacked up to form the information matrix. It is worth noting that the relations

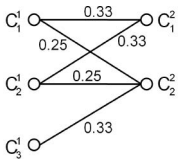


Fig. 3. An example of a cluster network, where each edge is marked with its weight.

between base clusterings' decisions have not been explored by these previous attempts. To this point, the link-based approach presented here is the first to address and make explicit use of this additional information. It aims to refine the BA matrix using the similarity within and between the input clustering solutions. This can be efficiently estimated from the link network representing the ensemble under examination. Having acquired such information, the refined cluster-association ($RA \in [0,1]^{N\times P}$) matrix is created by following the process summarized below.

Given a BA matrix, the entries representing "nil" associations (i.e., "0") are approximated from known ones (i.e., "1"), whose association degrees are preserved within the resulting RA matrix. In other words, $\forall x_i \in X, cl \in \Pi$, $BA(x_i, cl) = 1 \rightarrow RA(x_i, cl) = 1$. For each clustering π_g, $g = 1 \ldots M$, and their corresponding clusters $C_1^g, \ldots, C_{k_g}^g$, the association degree $RA(x_i, cl) \in [0,1]$ that data point $x_i \in X$ has with each cluster $cl \in \{C_1^g, \ldots, C_{k_g}^g\}$ is estimated by

$$RA(x_i, cl) = \begin{cases} 1, & \text{if } cl = C_*^g(x_i), \\ sim(cl, C_*^g(x_i)), & \text{otherwise}, \end{cases} \quad (1)$$

where $C_*^g(x_i)$ is a cluster label to which data point x_i has been assigned. In addition, $sim(C_x, C_y) \in [0,1]$ denotes the similarity between any two clusters $C_x, C_y \in \pi_g$, which can be discovered using the link-based similarity algorithms emphasized next.

3.2 Link-Based Similarity Algorithms

The RA matrix that is the main focus of link-based cluster ensembles is generated using the new algorithms presented in this section. Given a cluster ensemble Π of a set of data points X, a weighted graph $G = (V, W)$ can be constructed, where V is the set of vertices each representing a cluster in Π and W is a set of weighted edges between clusters. Formally, the weight $|w_{xy}| \in [0,1]$ assigned to the edge $w_{xy} \in W$ that connects vertices $v_x, v_y \in V$ (corresponding to clusters $C_x, C_y \in \Pi$) is estimated in accordance with the proportion of overlapping data members:

$$|w_{xy}| = \frac{|L_x \cap L_y|}{|L_x \cup L_y|}, \quad (2)$$

where $L_z \subset X$ denotes the set of data points belonging to cluster $C_z \in \Pi$. Note that G is an undirected graph such that $|w_{xy}| = |w_{yx}|, \forall v_x, v_y \in V$. Fig. 3 shows the network of clusters that is generated from the example given in Fig. 1. In particular, circle nodes represent clusters and edges existing only when the corresponding weights are nonzero. This graph represents the similarity among clusters which is gauged only by the basis of shared content. As such, the weight of edges between clusters of the same clusterings (e.g., C_1^1 and C_2^1) are simply nil.

Shared neighbors have been widely recognized as the basic evidence to justify the similarity among vertices in a link network [18], [33]. Formally, a vertex $v_z \in V$ is a common neighbor of vertices $v_x, v_y \in V$, provided that $|w_{xz}|, |w_{yz}| > 0$. Many advanced methods extend this node-based basis by taking into account the common neighbors that may be many edges away from the two under examination: for instance, SimRank [25], PageSim [35], and a variation of random walk algorithms [14], [37]. Despite reported effectiveness, these techniques are computationally expensive, or even impractical for a large data set. Henceforth, WCT, WTQ, and CSM algorithms are proposed as part of the current research for efficiently approximating the similarity between clusters in the aforementioned link network.

3.2.1 Weighted Connected-Triple (WCT)

First, WCT extends the Connected-Triple method [43] that has been developed to identify ambiguous author names within a publication database. The initial technique is built on a social network represented as an undirected graph $G' = (V, E)$, where V is the set of vertices each corresponding to an author name and E is the set of unweighted edges each standing for a coauthorship relation. With this network, the similarity of vertices $v_x, v_y \in V$ can be estimated by counting the number of Connected-Triples (i.e., triples) they are part of. Formally, a triple, $Triple = (V_{Triple}, E_{Triple})$, is a subgraph of G' containing three vertices $V_{Triple} = \{v_x, v_y, v_k\} \subset V$ and two edges $E_{Triple} = \{e_{xk}, e_{yk}\} \subset E$, with $e_{xy} \notin E$. This simple counting might be sufficient for any indivisible object, e.g., data point or author. However, to evaluate the similarity between clusters, it is important to realize and take into account the composite characteristic of a cluster, i.e., shared data members.

Inspired by this idea, the WCT algorithm is established. With a weighted graph $G = (V, W)$, presented in Fig. 3, the WCT measure of vertices $v_x, v_y \in V$ with respect to each center of a triple $v_z \in V$ is defined as

$$WCT_{xy}^z = \min(|w_{xz}|, |w_{yz}|), \quad (3)$$

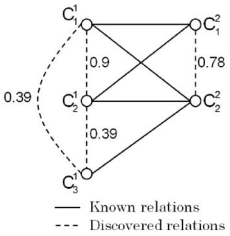
where $|w_{xz}|$ and $|w_{yz}|$ are weights of the edges $w_{xz}, w_{yz} \in W$ connecting vertices v_x and v_z and vertices v_y and v_z, respectively. The summation of all triples $(1 \ldots \lambda)$ between vertices v_x and v_y can be calculated by

$$WCT_{xy} = \sum_{z=1}^{\lambda} WCT_{xy}^z. \quad (4)$$

Following that, the similarity $S_{WCT}(v_x, v_y)$ between vertices v_x and v_y (or cluster C_x and C_y) is defined as

$$S_{WCT}(v_x, v_y) = \frac{WCT_{xy}}{WCT_{max}} \times DC, \quad (5)$$

where $DC \in [0,1]$ is a constant decay factor (i.e., confidence level of accepting two nonidentical clusters as being similar) and $WCT_{max} = \max_{\forall v_p, v_q \in V} WCT_{pq}$. With this link-based similarity metric, $S_{WCT}(v_x, v_y) \in [0,1]$ with $S_{WCT}(v_x, v_x) = 1$, $\forall v_x, v_y \in V$. It is also reflexive such that $S_{WCT}(v_x, v_y)$ is equivalent to $S_{WCT}(v_y, v_x)$. Following the example shown in



	C_1^1	C_2^1	C_3^1	C_1^2	C_2^2
x_1	1	0.9	0.39	1	0.78
x_2	1	0.9	0.39	0.78	1
x_3	0.9	1	0.39	1	0.78
x_4	0.9	1	0.39	0.78	1
x_5	0.39	0.39	1	0.78	1

RA matrix

Fig. 4. An example of discovered associations using the WCT algorithm and the resulting RA matrix, where $DC = 0.9$.

Figs. 1 and 3, the resulting RA matrix and those associations discovered by WCT are presented in Fig. 4.

3.2.2 Weighted Triple-Quality (WTQ)

Unlike WCT, which concentrates solely on the magnitude of common triples, the Weighted Triple-Quality algorithm aims to differentiate their significance and hence their contributions toward the underlying similarity measure. WTQ is inspired by the initial measure of [1], which evaluates the association between personal home pages. In particular, features of the compared pages p_a and p_b are used to estimate their similarity score, $score(p_a, p_b)$, as follows:

$$score(p_a, p_b) = \sum_{\forall z_c \in Z} \frac{1}{\log(frequency(z_c))}, \tag{6}$$

where Z denotes the set of features shared by home pages p_a and p_b, and $frequency(z_d)$ represents the number of times z_d appears in the studied set of pages. Note that the method gives high weights to rare features and low weights to features that are common to most of the pages.

For WTQ, (6) can be modified to discriminate the quality of shared triples between a pair of vertices in question. Specifically, the quality of each vertex is determined by the rarity of links connecting itself to other vertices in a network. With a weighted graph $G = (V, W)$, the WTQ measure of vertices $v_x, v_y \in V$ with respect to each center of a triple $v_z \in V$ is estimated as follows:

$$WTQ_{xy}^z = \frac{1}{\sum_{\forall v_t \in N_z} |w_{zt}|}. \tag{7}$$

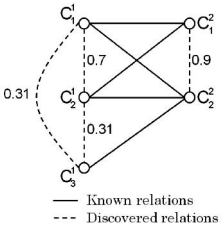
Here, $N_z \subset V$ denotes the set of vertices that is directly linked to the vertex v_z such that $\forall v_t \in N_z, |w_{zt}| > 0$. The accumulative WTQ score from all triples $(1 \dots \lambda)$ between vertices v_x and v_y can be approximated by

$$WTQ_{xy} = \sum_{z=1}^{\lambda} WTQ_{xy}^z. \tag{8}$$

Then, the similarity $S_{WTQ}(v_x, v_y)$ between vertices v_x and v_y is defined as

$$S_{WTQ}(v_x, v_y) = \frac{WTQ_{xy}}{WTQ_{max}} \times DC, \tag{9}$$

where WTQ_{max} is the maximum WTQ_{pq} value of any two vertices $v_p, v_q \in V$ and $DC \in [0, 1]$ is a constant decay factor. Note that the properties of the S_{WTQ} metric is similar to those of S_{WCT} previously emphasized. Based on the example shown in Figs. 1 and 3, the resulting RA matrix and those associations discovered by WTQ are presented in Fig. 5.



	C_1^1	C_2^1	C_3^1	C_1^2	C_2^2
x_1	1	0.7	0.31	1	0.9
x_2	1	0.7	0.31	0.9	1
x_3	0.7	1	0.31	1	0.9
x_4	0.7	1	0.31	0.9	1
x_5	0.31	0.31	1	0.9	1

RA matrix

Fig. 5. An example of discovered associations using the WTQ algorithm and the resulting RA matrix, where $DC = 0.9$.

3.2.3 Combined Similarity Measure (CSM)

With the objective of obtaining a robust similarity evaluation, this particular algorithm combines the WCT and WTQ measures previously described. From (3) and (7), the CSM measure between $v_x, v_y \in V$ with respect to each center of a triple $v_z \in V$ is

$$CSM_{xy}^z = \frac{\min(|w_{xz}|, |w_{yz}|)}{\sum_{\forall v_t \in N_z} |w_{zt}|}. \tag{10}$$

The accumulative CSM measure from all triples $(1 \dots \lambda)$ between v_x and v_y is approximated by

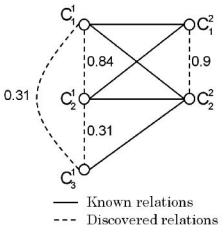
$$CSM_{xy} = \sum_{z=1}^{\lambda} CSM_{xy}^z. \tag{11}$$

The similarity $S_{CSM}(v_x, v_y)$ between vertices v_x and v_y is defined by the following:

$$S_{CSM}(v_x, v_y) = \frac{CSM_{xy}}{CSM_{max}} \times DC, \tag{12}$$

where CSM_{max} is the maximum CSM_{pq} value of any two vertices $v_p, v_q \in V$ and $DC \in [0, 1]$ is a constant decay factor. The core properties of the previous metrics are also held for S_{CSM}. By following the example given in Figs. 1 and 3, the resulting RA matrix and those associations discovered by CSM are presented in Fig. 6.

It is noteworthy that a similar graph representing the equivalent information to a BA matrix has also been employed by the Metaclustering Algorithm (MCLA) of [44]. However, that graph is only used to represent the content-based similarity among clusters, while the discovery of link-based relations has not been attempted. MCLA does not focus on these additional associations, but simply applies the METIS technique [27] to partition this graph into metaclusters from which the final data partition is



	C_1^1	C_2^1	C_3^1	C_1^2	C_2^2
x_1	1	0.84	0.31	1	0.9
x_2	1	0.84	0.31	0.9	1
x_3	0.84	1	0.31	1	0.9
x_4	0.84	1	0.31	0.9	1
x_5	0.31	0.31	1	0.9	1

RA matrix

Fig. 6. An example of discovered associations using the CSM algorithm and the resulting RM matrix, where $DC = 0.9$.

approximated. Another previous attempt to bring in the similarity between clusters and data points together into consideration has been reported in [12] as the framework of "Hybrid Bipartite Graph Formulation (HBGF)." Despite this useful idea, the final data partition is acquired by applying the spectral graph partitioning algorithm [39] to the bipartite graph that represents the information equivalent to the BA matrix. In contrast to these methods, the link-based approach takes a step further as it induces additional knowledge from the basic graph and uses it to refine the BA matrix.

3.3 Consensus Methods for the RA Matrix

Having obtained the RA matrix, two different types of consensus methods are proposed here to generate the final data partition: Feature-Based and Bipartite Graph Partitioning, respectively.

3.3.1 Feature-Based Partitioning

For the RA matrix of dimension $N \times P$, each of P columns that represents the association degree to a specific cluster can be regarded as a new feature describing N data points. As such, the RA matrix is virtually a high-level data matrix, to which any numerical clustering technique can be directly applied. For computational efficiency, two simple data partitioning algorithms of k-means [21] and PAM [28] are employed in the current research (brief details of these algorithms are given below). In addition to its simplicity, the FBP method is a unique combination of numerical ensemble-based matrix and clustering techniques. It is significantly different from the existing feature-based approach to cluster ensemble (e.g., [46] and [47]) that concentrates on the clustering of categorical data, i.e., cluster labels. It is also important to note that FBP is different from the Weighted Similarity Partitioning Algorithm (WSPA) of [7]. Despite the refined variations of the BA matrix employed by these methods being similarly considered as a high-level data matrix, the concepts behind the creation of such matrices are dissimilar. Yet, according to the details published in [7], WSPA does not make use of k-means or PAM as a consensus function. Instead, the cosine measure is used to estimate the pairwise similarity among data points, which is then transformed to a similarity graph to which a graph partitioning technique (e.g., METIS) is applied.

k-means (KM) is perhaps, the best known clustering technique that partitions data points into clusters. It first randomly selects (predefined) k data points as initial centroids, to which the remaining data points are assigned. Following that, the centroid of each cluster is updated as the mean of all points in that cluster. This process is iterated until no changes are made to the centroids (i.e., no reassignment of any point from one cluster to another). Note that the deterministic implementation of k-means, namely, "global k-means" [34], is exploited in this work. This clustering approach aims to overcome the well-known problem of k-means, that its performance greatly depends on the initial conditions. Global k-means is an incremental method with the basic idea of obtaining an optimal solution (of k clusters) from a series of local searches (from $k = 1 \ldots k-1$). In other words, the solution of the $(t+1)$th repetition (i.e., $t+1$ data clusters) can be estimated from the result of the tth iteration, where $t, t+1 \in \{1 \ldots, k\}$. The solution of the $(t+1)$th step is selected from N alternatives, each containing t cluster centers from the previous stage and one new cluster center that can be any data point in X. The collection of $t+1$ cluster centers that optimizes the objective function (of the classic k-means) is chosen and fetched as the input to the $(t+2)$th repetition.

Partitioning Around Medoids (PAM), also called k-medoids clustering, is a variation of k-means with the objective of minimizing the within-cluster variance, i.e., $Var(k) = \sum_{p=1}^{k} \sum_{x_i \in C_p} d(x_i, m_p)$, where m_p is the medoid of cluster C_p, k is the number of clusters, and $d(x_i, m_p)$ denotes the distance between the data point x_i and m_p. Principally, PAM first selects k data points arbitrarily as the medoids of k clusters. Each remaining data point is inserted into the cluster whose medoid is most similar to it. In each iteration, a new medoid is determined for each cluster by finding the data point with minimum total distance to all other points of the cluster. Following that, all data points are reassigned in accordance with the new set of medoids. The algorithm terminates when there is no change with $Var(k)$. The deterministic implementation of PAM is obtained from the MATLAB Library for Robust Analysis (LIBRA), http://wis.kuleuven.be/stat/robust/. It uses the heuristics of [45] to select the set of initial medoids, i.e., $m_1 \ldots m_k$. At first, m_1 is the data point that minimizes $\sum_{i=1 \ldots N} d(x_i, m_1)$. Then, m_2 is chosen such that $Var(k=2)$ is also minimized. This is repeated until all k medoids are obtained.

3.3.2 Bipartite Graph Partitioning

Unlike the previous method, where a specified clustering technique can be directly applied to the RA matrix, the BGP requires the underlying matrix to be initially transformed into a weighted bipartite graph $G = (V, W)$, where $V = V^X \cup V^C$ is a set of vertices representing both data points V^X and clusters V^C and W denotes a set of weighted edges that can be defined as follows:

- $|w_{xy}| = 0$ when vertices $v_x, v_y \in V^X$.
- $|w_{xy}| = 0$ when vertices $v_x, v_y \in V^C$.
- Otherwise, $|w_{xy}| = RA(v_x, v_y)$, when vertices $v_x \in V^X$ and $v_y \in V^C$. Note that the graph G is bidirectional such that $|w_{xy}| = |w_{yx}|$.

Having acquired this graph, a spectral graph partitioning method [39] is applied to generate a final data partition. Principally, given a graph $G = (V, W)$, SPEC first finds the K largest eigenvectors $u_1, \ldots, u_K$ of W, which are used to formed another matrix U (i.e., $U = [u_1, \ldots, u_K]$), whose rows are then normalized to have unit length. By considering the row of U as K-dimensional embedding of the graph vertices, SPEC applies k-means to these embedded points in order to acquire the final clustering result.

Note that the SPEC graph-partitioning technique has been similarly applied to cluster ensemble problems by Fern and Brodley [12]. However, this initial approach, called Hybrid Bipartite Graph Formulation, is based on the bipartite graph, which is transformed from the conventional BA matrix. It is also important to note that BGP is similar to the "Weighted Bipartite Partitioning Algorithm (WBPA)" introduced in [7]. They similarly extend HBGF by creating the weighted bipartite graph to which SPEC is applied.

TABLE 1
Description of Data Sets: Number of Data Points (N), Number of Attributes (D), Number of Classes (K), and Source

Dataset	N	D	K	Source
Synthetic Dataset:				
4-gaussian	100	12	4	[31]
2-banana	200	2	2	[23]
Complex Image	500	2	11	modified from [31]
3-ring	600	2	3	modified from [31]
5-gaussian	600	2	5	modified from [31]
Real Dataset:				
Iris	150	4	3	UCI [2]
Wine	178	13	3	UCI [2]
Glass	214	9	6	UCI [2]
Ionosphere	351	34	2	UCI [2]
Diabetes	768	8	2	UCI [2]

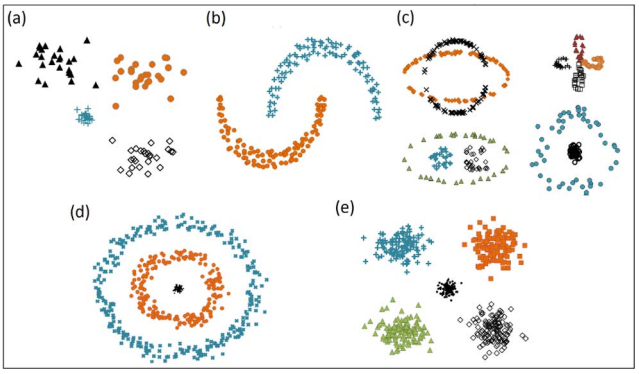


Fig. 7. Synthetic data sets: (a) 4-gaussian, (b) 2-banana, (c) Complex Image, (d) 3-ring, and (e) 5-gaussian.

However, the information used for the generation of such a graph is obtained differently. While WBPA employs the weighted distance metric to enrich the conventional crisp BA matrix, BGP accomplishes this using the similarity between clusters.

4 Performance Evaluation

This section presents the performance evaluation of the proposed link-based approach, using a number of benchmark validity criteria and data sets.

4.1 Investigated Data Sets

The experimental evaluation is conducted on 10 data sets. Table 1 summarizes the details of these data sets that are grouped into synthetic and real categories. Five synthetic data sets of 4-gaussian, 2-banana, Complex Image, 3-ring, and 5-gaussian are shown in Figs. 7a to 7e, respectively. Note that the first synthetic data set acquired from [31] is initially created in two dimensions and later added with 10 more dimensions of noise. In addition to the synthetic data collection, five real data sets obtained from the benchmark UCI repository [2] are also employed.

4.2 Experiment Design

Based on the proposed link-based framework, nine cluster ensemble methods can be established as different combinations of link-based similarity measures (WCT, WTQ, or CSM) used to create the RA matrix and a consensus function (FBP(KM), FBP(PAM), or BGP(SPEC)). Effectively, these techniques are assessed against several cluster ensemble algorithms found in the literature. Details of these compared methods and the experimental setting are exhibited below.

4.2.1 Compared Methods

In order to properly examine the potential or otherwise of the proposed methods, they are evaluated against baseline models and several state-of-the-art techniques developed for cluster ensembles.

Baseline methods. Unlike link-based methods, three distinct baseline techniques can be formulated as the combinations of the BA matrix and the consensus functions of FBP(KM), FBP(PAM), and BGP(SPEC), respectively. By including these models in the evaluation, it is possible to directly differentiate the quality of BA and RA matrices.

Pairwise similarity algorithms. This specific category of cluster ensemble method is based principally on the pairwise similarity among data points. Given an ensemble $\Pi = \{\pi_1, \ldots, \pi_M\}$, an $N \times N$ similarity matrix is constructed for each ensemble member, denoted as S_m, $m = 1 \ldots M$. An entry $S_m(x_i, x_j)$ that represents the similarity between two data points x_i and x_j in the mth ensemble member is 1 if $C(x_i) = C(x_j)$, and 0 otherwise. Following that, M similarity matrices are merged to form a co-association (CO) matrix [17]. Formally, the similarity between two data points x_i and x_j is summarized across all base clusterings as $CO(x_i, x_j) = \frac{1}{M}\sum_{m=1}^{M} S_m(x_i, x_j)$. Since the CO matrix is a similarity matrix, any similarity-based clustering algorithm can be applied to this matrix to yield the final partition. Among several existing similarity-based methods, the most well-known technique is the agglomerative clustering algorithm. Specific to [17], the single-linkage (SL) and average-linkage (AL) agglomerative clusterings are used to derive the final solution.

Graph-based algorithms. The graph-based ensemble methods of [44] (CSPA, HGPA, and MCLA), [12] (HBGF), and [7] (WSPA and WBPA) are employed in this evaluation. First, the Cluster-based Similarity Partitioning Algorithm (CSPA) creates a similarity graph, where vertices represent data points and edges' weight represent similarity scores obtained from the CO matrix. Afterward, a graph partitioning algorithm called METIS [27] is used to partition the similarity graph into k clusters. Based on the BA matrix, the Hyper-Graph Partitioning Algorithm (HGPA) constructs a hypergraph, where vertices represent data points and the same-weighted hyperedges represent clusters in the ensemble. Then, HMETIS [26] is applied to partition the underlying hypergraph into k parts with roughly of the same size. In addition, Meta-Clustering Algorithm (MCLA) generates a graph that represents the relationships among clusters in the ensemble. METIS is also employed to partition the metalevel graph into K metaclusters. Effectively, each data point has a specific association degree to each metacluster. This can be estimated from the number of original clusters to which the data point belongs in the metacluster. The final clustering is produced by assigning each data point to the metacluster with which it is most frequently associated.

Unlike the previous methods, Hybrid Bipartite Graph Formulation (HBGF) makes use of the bipartite graph whose vertices represent both data points and clusters. An edge between any data point and cluster is either 1 (when the data point belongs to the cluster) or 0 (otherwise). The spectral graph partitioning algorithm of [39] is exploited to obtain the final clustering. It is noteworthy that HBGF is identical to one of the aforementioned baseline models where the BGP(SPEC) consensus is exploited with the original BA matrix. To consolidate the evaluation, two graph-based techniques of [7], namely, Weighted Similarity Partitioning Algorithm (WSPA) and Weighted Bipartite Partitioning Algorithm (WBPA), are also included in this experiment. These advanced methods extend CSPA and HBGF, respectively, by refining the corresponding graphs with additional information obtained from the coupling of weighted distance metric and the soft subspace clustering solutions of Locally Adaptive Clustering (LAC) [8]. Specific to this experiment, METIS and SPEC are used as consensus functions for WSPA and WBPA, respectively. Please refer to the abovementioned publication for further details.

Feature-based algorithm. The method included in this evaluation was recently introduced in [40], as the Iterative Voting Consensus (IVC) algorithm. It aims to obtain the consensus partition π^* of the data set X from the categorical data induced by a cluster ensemble $\Pi = \{\pi_1, \ldots, \pi_M\}$. Principally, it utilizes the set of feature vectors $Y = \{y_1, \ldots, y_N\}$, with y_i, $i = 1 \ldots N$, being specified as $y_i = \{\pi_1(x_i), \ldots, \pi_M(x_i)\}$, where $\pi_g(x_i)$ represents a label of specific cluster in clustering π_g, $g = 1 \ldots M$, to which a data point x_i belongs. In each iteration, IVC first estimates the center of each cluster in π^*. Note that each cluster C_j, $j = 1 \ldots k$, in the target clustering π^* has a cluster center $center_j = \{mode(X_j, \pi_1), \ldots, mode(X_j, \pi_M)\}$, where $X_j \subset X$ is the set of data points belonging to the cluster C_j and $mode(X_j, \pi_g)$ denotes the majority labels (in the clustering π_g) of members of X_j. Having obtained these centers, IVC then reassigns each data point to its closest cluster center. This is possible using the Hamming distance between M-dimensional vectors that represent data points and cluster centers. The iterative process continues until there is no change with the target clustering π^*.

4.2.2 Experimental Setting

For comparison, as in [12], [17], and [19], each clustering method divides data points into a partition of K (the number of *true classes* for each data set) clusters, which is then evaluated against the corresponding true partition using the evaluation indices of: Normalized Mutual Information (NMI) [44], Classification Accuracy (CA) [40], and Rand Index (RI) [42]. Details of these quality measures are included in Section 1 of online supplemental material.[1] Note that true classes are known for all data sets but are not used by the cluster ensemble process. They are only used to evaluate the quality of the clustering results. Other specific settings of cluster ensembles are listed as follows:

- k-means is used to generate base clusterings, each with a random initialization of cluster centers. However, for WSPA and WBPA, the underlying ensemble is obtained from the repeated applications of LAC, each with a random initialization of cluster centers and the parameter h is arbitrarily selected from the range of $[0.25, 1]$.
- Three schemes for selecting the number of clusters (k) in each base clustering are: 1) True-k, where k equals the number of known classes (i.e., K), 2) Fixed-k where k is fixed to $\lceil\sqrt{N}\rceil$, and 3) Random-k where k is a random number in the range of $\{2, \lceil\sqrt{N}\rceil\}$. The last two strategies aim to generate diversity in the ensemble by following the intuition introduced by Fred and Jain [17], Hadjitodorov et al. [20], Kuncheva and Vetrov [31]. It is suggested that k should be greater than the expected number of clusters and the common rule of thumb is $k = \sqrt{N}$.
- An ensemble size (M) of 10 is used for experimentation.
- The decay factor (DC) of 0.9 is used with the three link-based similarity algorithms.
- The quality of each cluster ensemble method with respect to a specific ensemble setting is generalized as the average of 50 runs.

4.3 Experiment Results

Based on the NMI measure, Tables 2 and 3 compare the performance of different cluster ensemble methods over synthetic and real data sets, respectively. Note that Max(LAC), Avg(LAC), and Min(LAC) represent the maximum, average and minimum NMI scores among base clusterings (i.e., LAC) of the ensembles used by WSPA and WBPA. Likewise, Max(base), Avg(base), and Min(base) denote the similar statistics between input clusterings (i.e., k-means) of the ensembles employed by other methods. It is clear that the link-based techniques usually generate data partitions of higher quality than their baseline models and other compared methods. Also, in several data sets, their NMI measures exceed the maximum of the corresponding base clusterings, i.e., Max(base). Note that the three graph-based methods (CSPA, HGPA, and MCLA) and pairwise-similarity algorithms ($\mathrm{CO} + \mathrm{SL}$ and $\mathrm{CO} + \mathrm{AL}$) are rather effective over synthetic data sets, but not with real data. While it is generally less accurate than the link-based methods, WBPA is exceptionally accurate for a few data sets such as Wine. In the presence of noise such as the 4-gaussian data set, both WSPA and WBPA are inaccurate due to the fact that LAC provides low quality base clusterings. Similar experimental results with these methods are observed using CA and RI evaluation indices. See further details in Table 4 and Section 2 of the online supplemental material.

In order to further evaluate the quality of identified techniques, the number of times that one method is significantly *better* and *worse* (of 95 percent confidence level) than the others are assessed across experimented data sets. Let $\overline{X}_C(i, \beta)$ be the average value of validity index $C \in \{CA, NMI, RI\}$ across n runs ($n = 50$ in this evaluation) for a clustering method $i \in CM$ (CM is a set of 20 experimented clustering methods), on a specific experiment setting $\beta \in ST$ (ST is a set of 30 unique combination of three ensemble types and 10 data sets). The 95 percent confidence interval, $[L_{\overline{X}_C(i,\beta)}, U_{\overline{X}_C(i,\beta)}]$, for the mean $\overline{X}_C(i, \beta)$ of validity criterion C is defined by

1. Available at http://itschool.mfu.ac.th/~natthakan/tpami2011/.

TABLE 2
NMI Measures of Different Cluster Ensemble Methods on Five Synthetic Data Sets

Method	4-gaussian			2-banana			Complex Image			3-ring			5-gaussian		
	True-k	Fixed-k	Random-k	True-k	Fixed-k	Random-k	True-k	Fixed-k	Random-k	True-k	Fixed-k	Random-k	True-k	Fixed-k	Random-k
WCT+KM	0.942	**0.970**	**0.951**	0.344	0.976	0.808	0.734	0.748	0.738	0.151	0.965	0.331	**1.000**	**1.000**	**1.000**
WCT+PAM	0.941	**0.970**	**0.952**	0.344	0.750	0.737	0.723	0.715	0.720	0.146	0.700	0.291	**1.000**	**1.000**	**1.000**
WCT+SPEC	0.941	0.969	0.949	0.344	**1.000**	0.788	0.722	0.735	0.725	0.142	**1.000**	0.264	**1.000**	**1.000**	**1.000**
WTQ+KM	**0.943**	0.968	**0.951**	0.344	**1.000**	**0.923**	**0.736**	**0.759**	**0.744**	**0.174**	0.981	**0.345**	**1.000**	**1.000**	0.996
WTQ+PAM	**0.943**	**0.970**	**0.951**	0.344	0.900	0.874	0.733	0.725	0.735	0.157	0.709	0.329	**1.000**	**1.000**	**1.000**
WTQ+SPEC	**0.943**	0.967	**0.951**	0.344	**1.000**	0.903	**0.736**	0.735	0.740	0.145	0.995	0.217	**1.000**	**1.000**	**1.000**
CSM+KM	0.942	0.966	**0.951**	0.344	0.980	0.833	0.708	**0.751**	**0.741**	0.149	0.967	0.326	**1.000**	**1.000**	**1.000**
CSM+PAM	0.942	**0.970**	**0.951**	0.344	0.757	0.819	0.719	0.719	0.724	0.145	0.740	0.325	**1.000**	**1.000**	**1.000**
CSM+SPEC	0.941	0.967	0.950	0.344	**1.000**	0.827	0.728	0.739	0.735	0.141	**1.000**	0.282	**1.000**	**1.000**	**1.000**
BM+KM	0.941	0.956	0.942	0.344	0.527	0.528	0.734	0.734	0.740	**0.199**	0.724	0.290	**1.000**	0.994	0.998
BM+PAM	0.938	0.919	0.938	0.344	0.170	0.443	0.734	0.684	0.726	0.159	0.224	0.163	**1.000**	0.592	0.977
HBGF	0.939	0.965	0.941	0.344	0.801	0.521	0.669	0.631	0.692	0.150	0.730	0.235	0.985	0.765	0.975
CSPA	**0.950**	**0.970**	**0.960**	**0.354**	**1.000**	0.588	0.667	0.675	0.670	0.157	0.736	0.237	0.999	**1.000**	0.998
HGPA	0.462	0.960	0.939	0.000	**1.000**	0.855	0.655	0.706	0.665	0.042	0.745	**0.412**	0.096	**1.000**	**1.000**
MCLA	0.939	0.968	**0.951**	0.346	**1.000**	**0.959**	0.708	0.701	0.702	0.142	0.731	0.296	0.999	**1.000**	0.996
CO+SL	0.925	0.905	0.932	0.346	**1.000**	**0.968**	**0.782**	**0.771**	**0.781**	**0.477**	0.992	**0.719**	0.995	**1.000**	0.997
CO+AL	0.940	0.960	0.943	0.346	**1.000**	0.647	0.701	0.698	0.690	0.143	**0.997**	0.204	0.999	**1.000**	0.987
IVC	0.804	0.807	0.834	0.229	0.162	0.347	0.705	0.649	0.700	0.160	0.279	0.184	0.833	0.602	0.781
WSPA	0.284	0.299	0.265	**0.443**	0.446	0.444	0.620	0.617	0.619	0.116	0.198	0.168	0.998	**1.000**	0.999
WBPA	0.207	0.213	0.208	**0.457**	0.445	0.447	0.709	0.716	0.723	0.133	0.145	0.140	**1.000**	**1.000**	**1.000**
Max(base)	0.939	0.789	0.910	0.356	0.518	0.614	0.772	0.730	0.775	0.156	0.573	0.606	1.000	0.734	0.927
Avg(base)	0.873	0.763	0.799	0.347	0.511	0.531	0.700	0.704	0.696	0.142	0.566	0.470	0.906	0.722	0.793
Min(base)	0.722	0.737	0.667	0.343	0.508	0.401	0.636	0.670	0.605	0.131	0.562	0.180	0.808	0.715	0.671
Max(LAC)	0.292	0.332	0.321	0.457	0.520	0.592	0.767	0.732	0.791	0.155	0.574	0.593	1.000	0.735	0.962
Avg(LAC)	0.185	0.264	0.192	0.457	0.513	0.530	0.702	0.706	0.712	0.142	0.566	0.423	0.906	0.722	0.843
Min(LAC)	0.076	0.199	0.046	0.457	0.508	0.454	0.645	0.676	0.637	0.132	0.561	0.245	0.813	0.714	0.716

The three highest NMI scores of each experimental setting are highlighted in **boldface**. *Note that the last six rows are the statistics of base clusterings.*

TABLE 3
NMI Measures of Different Cluster Ensemble Methods on Five Real Data Sets

Method	Iris			Wine			Glass			Ionosphere			Diabetes		
	True-k	Fixed-k	Random-k	True-k	Fixed-k	Random-k	True-k	Fixed-k	Random-k	True-k	Fixed-k	Random-k	True-k	Fixed-k	Random-k
WCT+KM	0.655	0.769	0.816	**0.881**	**0.807**	0.838	0.318	**0.388**	0.312	0.134	0.112	0.123	0.101	**0.117**	0.052
WCT+PAM	0.656	0.810	0.802	**0.881**	0.805	0.843	0.311	0.343	0.324	0.134	**0.228**	**0.170**	0.096	0.055	0.046
WCT+SPEC	0.658	**0.851**	**0.824**	0.871	0.800	**0.845**	**0.336**	0.360	0.328	0.134	0.117	0.133	0.093	0.090	0.046
WTQ+KM	0.662	0.706	0.789	0.867	0.771	0.791	0.320	0.375	0.315	0.134	0.123	0.126	0.078	0.075	0.054
WTQ+PAM	**0.674**	0.774	0.772	0.856	0.793	0.826	0.307	0.354	0.328	0.134	**0.145**	0.130	0.079	0.048	0.050
WTQ+SPEC	0.610	0.829	0.799	0.842	0.788	0.829	0.326	0.372	**0.352**	0.134	0.133	0.128	0.080	0.060	0.053
CSM+KM	0.660	0.740	0.812	0.869	0.792	0.830	0.317	**0.380**	0.316	**0.135**	0.124	0.127	0.104	**0.114**	0.049
CSM+PAM	0.658	0.799	0.796	0.866	0.798	0.840	0.310	0.351	0.329	**0.135**	**0.239**	**0.171**	0.098	0.054	0.048
CSM+SPEC	0.657	0.843	**0.818**	0.870	0.795	0.844	**0.338**	0.358	0.334	**0.135**	0.143	**0.137**	0.097	0.089	0.048
BM+KM	0.652	0.785	0.747	**0.875**	**0.810**	0.727	0.323	0.281	0.318	0.134	0.126	0.108	0.103	0.109	**0.097**
BM+PAM	0.651	0.416	0.706	0.874	0.483	0.831	0.298	0.308	0.339	0.134	0.063	0.134	0.099	0.041	0.041
HBGF	0.650	0.677	0.730	0.871	0.793	**0.859**	0.328	0.363	0.338	0.134	0.111	0.116	0.100	0.004	0.029
CSPA	**0.666**	**0.857**	**0.830**	0.816	0.782	0.781	0.243	0.301	0.282	0.100	0.089	0.108	**0.110**	0.057	**0.086**
HGPA	0.274	**0.850**	0.805	0.246	0.802	0.818	0.235	0.278	0.299	0.030	0.106	0.070	0.000	0.040	0.020
MCLA	0.657	0.802	0.810	0.871	0.767	0.810	0.280	0.326	0.324	**0.135**	0.134	0.117	0.099	0.042	0.046
CO+SL	0.660	0.722	0.728	0.846	0.384	0.419	0.308	**0.410**	**0.340**	**0.135**	0.075	0.048	0.037	0.018	0.006
CO+AL	0.654	0.765	0.757	0.874	0.735	0.836	0.319	0.364	0.323	**0.135**	0.130	0.121	0.082	0.004	0.022
IVC	0.627	0.356	0.606	0.778	0.568	0.751	0.221	0.281	0.303	0.113	0.061	0.102	0.072	0.034	0.043
WSPA	**0.669**	0.662	0.667	0.725	0.645	0.721	0.324	0.310	0.294	0.120	0.128	0.131	**0.116**	**0.113**	0.065
WBPA	0.642	0.621	0.640	0.866	**0.881**	**0.883**	**0.345**	0.344	**0.346**	0.126	0.126	0.121	**0.116**	0.085	**0.071**
Max(base)	0.669	0.632	0.742	0.892	0.618	0.831	0.382	0.428	0.412	0.135	0.296	0.303	0.124	0.113	0.109
Avg(base)	0.644	0.606	0.657	0.856	0.592	0.662	0.318	0.354	0.349	0.132	0.244	0.230	0.079	0.104	0.093
Min(base)	0.598	0.580	0.596	0.742	0.566	0.542	0.257	0.281	0.260	0.110	0.217	0.084	0.040	0.095	0.067
Max(LAC)	0.669	0.614	0.659	0.891	0.620	0.874	0.380	0.428	0.408	0.130	0.337	0.321	0.128	0.114	0.107
Avg(LAC)	0.644	0.579	0.606	0.859	0.585	0.724	0.320	0.391	0.316	0.124	0.301	0.248	0.080	0.103	0.091
Min(LAC)	0.595	0.536	0.547	0.769	0.549	0.575	0.253	0.354	0.176	0.098	0.261	0.141	0.033	0.093	0.073

The three highest NMI scores of each experimental setting are highlighted in **boldface**. *Note that the last six rows are the statistics of base clusterings.*

$$L_{\overline{X}_C(i,\beta)} = \overline{X}_C(i,\beta) - 1.96\frac{S_C(i,\beta)}{\sqrt{n}} \text{ and}$$

$$U_{\overline{X}_C(i,\beta)} = \overline{X}_C(i,\beta) + 1.96\frac{S_C(i,\beta)}{\sqrt{n}}.$$

Note that $S_C(i,\beta)$ is the standard deviation of the validity index C across n runs for a clustering method i and an experiment setting β. The number of times that one method $i \in CM$ is significantly *better* than others, $B_C(i)$ (in accordance with the validity criterion C), can be estimated by

TABLE 4
CA Measures of Different Cluster Ensemble Methods on Five Real Data Sets

Method	Iris			Wine			Glass			Ionosphere			Diabetes		
	True-k	Fixed-k	Random-k	True-k	Fixed-k	Random-k	True-k	Fixed-k	Random-k	True-k	Fixed-k	Random-k	True-k	Fixed-k	Random-k
WCT+KM	0.827	0.909	0.913	**0.968**	**0.934**	0.948	0.549	**0.619**	0.529	0.711	0.663	0.699	0.690	0.652	0.653
WCT+PAM	0.829	0.928	0.900	**0.968**	**0.934**	0.954	0.551	0.573	0.563	0.711	**0.776**	**0.740**	**0.692**	0.655	0.658
WCT+SPEC	0.819	**0.951**	**0.920**	0.964	0.933	**0.955**	**0.559**	0.580	0.543	**0.712**	0.688	0.709	0.688	0.653	0.652
WTQ+KM	0.834	0.874	0.892	0.962	0.906	0.927	0.556	**0.614**	0.555	0.711	0.695	0.702	0.676	0.655	0.654
WTQ+PAM	**0.845**	0.890	0.871	0.957	0.928	0.948	0.557	0.588	**0.607**	0.711	**0.721**	0.706	0.680	**0.658**	**0.662**
WTQ+SPEC	0.806	0.936	0.902	0.949	0.926	0.948	**0.560**	0.610	**0.601**	**0.712**	0.708	0.704	0.675	0.655	0.653
CSM+KM	0.832	0.892	0.909	0.963	0.922	0.941	0.551	**0.612**	0.530	**0.712**	0.691	0.704	**0.691**	0.652	0.652
CSM+PAM	0.828	0.920	0.897	0.962	0.932	0.953	0.557	0.578	0.571	**0.712**	**0.784**	**0.740**	**0.693**	0.656	0.658
CSM+SPEC	0.821	0.946	0.916	0.963	0.931	0.954	0.558	0.580	0.543	**0.712**	0.714	**0.713**	0.689	0.654	0.652
BM+KM	0.826	0.911	0.863	**0.966**	0.932	0.848	0.550	0.545	0.589	0.711	0.655	0.679	0.686	0.651	**0.661**
BM+PAM	0.825	0.678	0.840	0.965	0.722	0.940	0.550	0.604	**0.623**	0.711	0.645	0.703	0.688	0.651	0.655
HBGF	0.824	0.796	0.848	0.964	0.922	**0.961**	0.554	0.532	0.599	**0.712**	0.683	0.695	0.684	0.651	0.655
CSPA	**0.841**	**0.956**	**0.934**	0.935	0.923	0.924	0.549	0.597	0.581	0.674	0.665	0.680	0.679	0.652	**0.668**
HGPA	0.587	**0.951**	**0.920**	0.559	0.932	0.942	0.525	0.544	0.563	0.644	0.677	0.660	0.651	0.652	0.652
MCLA	0.832	0.916	0.918	0.964	0.918	0.939	0.541	0.589	0.576	**0.712**	0.698	0.691	0.691	0.652	0.653
CO+SL	0.805	0.760	0.783	0.923	0.553	0.579	0.515	0.506	0.489	**0.712**	0.646	0.643	0.664	0.655	0.652
CO+AL	0.824	0.838	0.852	0.965	0.859	0.949	0.550	0.511	0.532	**0.712**	0.698	0.694	0.683	0.651	0.655
IVC	0.762	0.623	0.707	0.843	0.735	0.874	0.487	0.536	0.561	0.701	0.662	0.691	0.682	0.655	0.656
WSPA	**0.843**	0.836	0.841	0.904	0.869	0.902	**0.611**	0.592	0.583	0.689	0.696	0.698	0.685	**0.683**	0.655
WBPA	0.838	0.818	0.836	0.962	**0.970**	**0.968**	0.553	0.544	0.552	0.701	0.705	0.697	0.690	**0.671**	0.653
Max(base)	0.852	0.968	0.966	0.972	0.971	0.971	0.578	0.697	0.680	0.712	0.930	0.917	0.710	0.766	0.755
Avg(base)	0.795	0.956	0.904	0.951	0.953	0.926	0.545	0.656	0.577	0.710	0.911	0.869	0.682	0.751	0.721
Min(base)	0.681	0.933	0.742	0.856	0.930	0.753	0.496	0.585	0.466	0.695	0.892	0.774	0.659	0.736	0.676
Max(LAC)	0.853	0.963	0.947	0.971	0.972	0.972	0.577	0.705	0.678	0.709	0.926	0.910	0.713	0.763	0.752
Avg(LAC)	0.799	0.934	0.855	0.954	0.947	0.949	0.546	0.665	0.541	0.706	0.902	0.842	0.683	0.748	0.711
Min(LAC)	0.675	0.887	0.750	0.878	0.917	0.911	0.498	0.605	0.430	0.689	0.874	0.741	0.657	0.732	0.677

The three highest CA scores of each experimental setting are highlighted in **boldface**. *Note that the last six rows are the statistics of base clusterings.*

$$B_C(i) = \sum_{\forall \beta \in ST} \sum_{\forall i^* \in CM, i^* \neq i} better_C^{\beta}(i, i^*), \tag{13}$$

$$better_C^{\beta}(i, i^*) = \begin{cases} 1, & \text{if } L_{\overline{X}_C(i,\beta)} > U_{\overline{X}_C(i^*,\beta)}, \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

Likewise, the number of times that one method $i \in CM$ is significantly *worse* than its competitors, $W_C(i)$, with respect to the validity index C is defined as

$$W_C(i) = \sum_{\forall \beta \in ST} \sum_{\forall i^* \in CM, i^* \neq i} worse_C^{\beta}(i, i^*), \tag{15}$$

$$worse_C^{\beta}(i, i^*) = \begin{cases} 1, & \text{if } U_{\overline{X}_C(i,\beta)} < L_{\overline{X}_C(i^*,\beta)}, \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

Using the aforementioned assessment formalism, Fig. 8 summarizes for each method $i \in CM$ the statistics of total performance $(B - W)_i = \sum_{\forall C \in \{CA, NMI, RI\}} B_C(i) - W_C(i)$. The results shown in this figure indicate the superior effectiveness of the proposed link-based methods as compared to other cluster ensemble techniques included in this experiment. In addition, CSPA and IVC appear to be the most and the least accurate among the compared methods, respectively. See Section 2 of the online supplemental material for the detailed statistics, which are categorized in accordance with evaluation indices and ensemble types.

Another important investigation is on the subject of relations between performance of cluster ensemble methods and different ensemble types. Fig. 9 presents the validity scores (summarized across all validity indices and data sets) of the three WCT-based techniques and "Others" that is the average performance of all compared methods. It is clearly shown for all ensemble types that the linked-based models are consistently more effective than the "Others." Also, all the examined methods, including the proposed ones, are

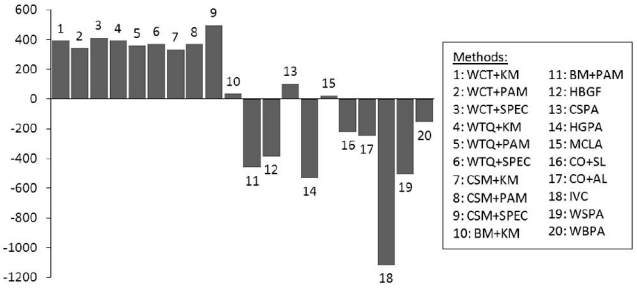


Fig. 8. The statistics of total performance, summarized across all evaluation indices, i.e., $(B - W)_i$, $\forall i \in CM$.

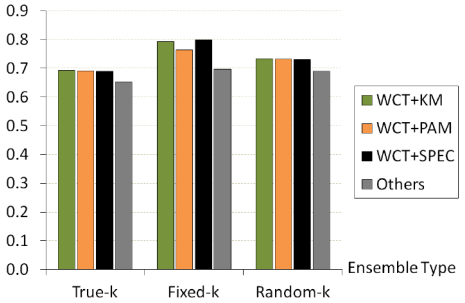


Fig. 9. Average validity score of each cluster ensemble method, summarized across all validity indices and data sets. These statistics are presented by the ensemble type, with "Others" denoting the average performance of all compared methods.

TABLE 5
Runtime (in Seconds) Required by Link-Based Similarity Algorithms to Construct the RA Matrix

Dataset	N	P	Run time (in seconds)		
			WCT	WTQ	CSM
4-gaussian	100	100	0.016	0.031	0.031
2-banana	200	150	0.063	0.078	0.078
Complex Image	500	230	0.250	0.266	0.266
3-ring	600	250	0.328	0.359	0.359
5-gaussian	600	250	0.328	0.359	0.359
Iris	150	130	0.047	0.063	0.063
Wine	178	140	0.063	0.073	0.073
Glass	214	150	0.063	0.078	0.078
Ionosphere	351	190	0.141	0.156	0.156
Diabetes	768	280	0.500	0.516	0.516

more accurate using the ensemble type of Fixed-k or Random-k, as compared to the True-k. A similar trend is also observed with both the WTQ and CSM-based methods. See these results in Section 2 of the online supplemental material.

4.4 Time Complexity and Parameter Analysis

Besides the previous quality assessments, computational time requirements of the link-based methods are discussed here. For the WCT algorithm, the time complexity of creating the RA matrix is $O(P^2l + NP)$, where N is the number of data points, P denotes the number of all clusters in an ensemble Π, and l represents the average number of neighbors connecting to one cluster in a link network of clusters. For each entry (corresponding to clusters $C_x, C_y \in \Pi$) in the $P \times P$ matrix of cluster similarity, WCT searches through l neighbors of C_x (or C_y) to identify triples. Following this, the RA matrix of size $N \times P$ is created using the aforementioned similarity matrix.

As the extension of WCT, WTQ continues searching through l neighbors of each potential triple identified earlier. Hence, the time complexity of WTQ is $O(P^2l^2 + NP)$, which is the same for CSM. For each of these link-based similarity algorithms, Table 5 presents the actual computational time (in seconds) required for creating the RA matrices from the experimented data sets, using the "Fixed-k" ensemble generation scheme. These algorithms were implemented in MATLAB and all experiments were conducted on a workstation with Intel(R)-Core(TM)2 CPU@2.40 GHz and 2 GB RAM. The execution times have been measured using the *cputime()* MATLAB function. Also, please consult Section 3 of online supplemental material for the corresponding scalability test.

In addition, the parameter analysis is also conducted with the aim of providing a practical means by which users can make the best use of the proposed link-based framework. Essentially, the performance of the resulting techniques are dependent to the DC value, which is used in estimating the similarity among clusters and refining the original BA matrix. To this extent, Fig. 10 illustrates such a relationship, based on the average of three validity measures (CA, NMI, and RI) across all data sets used in the experiments. For all link-based similarity algorithms of WCT, WTQ, and CSM, high DC values (i.e., 0.7 to 0.9) bring about a data partition of exceptionally good quality, as compared to those generated by other cluster ensemble methods (whose average validity scores are presented as *Others* in Fig. 10). Another important observation is that the effectiveness of link-based measures decreases as DC becomes smaller. Intuitively, the significance of disclosed memberships becomes trivial when DC is low. Hence, they may be overlooked by a consensus function and the quality of the resulting data partition is not improved. It is also noteworthy that the evaluation scores of CSM is superior than the maximum measure among base clusterings (i.e., $Max(base)$), even when DC is around 0.3 to 0.5. This suggests that CSM is more robust than the other link-based algorithms.

4.5 Diversity versus Accuracy Analysis

Existing research on cluster ensembles has pointed out that the diversity among ensemble members is a crucial factor for the quality of cluster ensembles [11], [20], [30]. It is shown in [11] that the high diversity among input data partitions correlates to the high quality of the final solution. This relation between diversity and accuracy is also extensively discussed by Kuncheva and Hadjitodorov [30] and Hadjitodorov et al. [20], with the former suggesting that a more accurate partition can be obtained from a diverse ensemble as compared to the nondiverse case. Specific to [20], it has been concluded, based on experimental studies, that medium diversity within the ensemble is preferred for the high quality of the final clustering result. Following the studies discussed above, the issue of diversity and accuracy of cluster ensembles is investigated here for a better understanding on the behaviors of the link-based methods. The corresponding findings may provide useful guidance for the creation of ensembles that promote the effectiveness of the proposed approach.

There are several different measures of diversity proposed in the literature of cluster ensembles. The majority of them are based on the matching of labels acquired from two data partitions. The present investigation follows the framework introduced by Fern and Brodley [11], which is based on the "pairwise" measurement of NMI among base

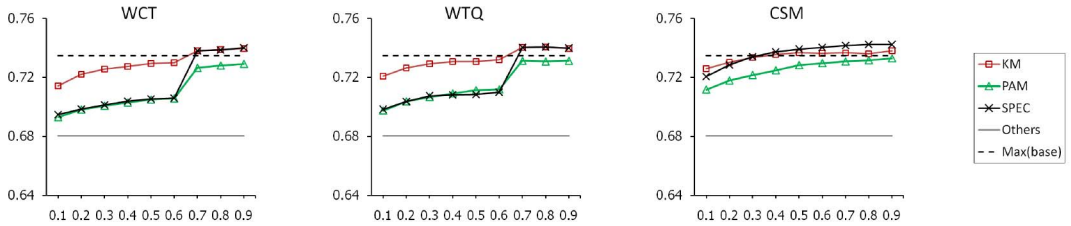


Fig. 10. The relations between $DC \in \{0.1, 0.2, \ldots, 0.9\}$ and the performance of link-based algorithms (the average of CA, NMI, and RI measures), whose values are presented in the X-axis and Y-axis, respectively. Note that $Others$ and $Max(base)$ denote the average evaluation measure across all compared cluster ensemble methods and the maximum measure among ensemble members, respectively.

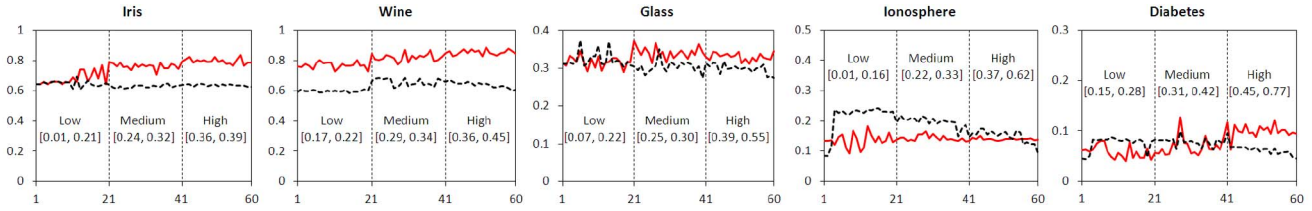


Fig. 11. The accuracy of the link-based approach $ACC(\pi^*)$ (presented by the red line) and the base clusterings $ACC(\Pi)$ (presented by the dashed black line), with respect to three levels of diversity for the Iris, Wine, Glass, Ionosphere, and Diabetes data sets.

clustering solutions. In particular, the diversity $DS(\Pi)$ of a given ensemble Π of size M is the average of all pairwise diversities among members of Π:

$$DS(\Pi) = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} (1 - NMI(\pi_i, \pi_j)), \qquad (17)$$

where $(1 - NMI(\pi_i, \pi_j))$ denotes the diversity between base clusterings $\pi_i, \pi_j \in \Pi$. Also, the higher $DS(\Pi)$ is, the more diverse the ensemble becomes.

To find out the preferred level of diversity (i.e., low, medium, or high) for link-based methods, the following controlled experiment is performed.

1. Generate ensembles of size 20 using each of three ensemble types (True-k, Fixed-k, and Random-k). Merge all these base partitions to form a pool of different clustering solutions (of size 60).
2. Randomly select one partition from this pool.
3. Construct an ensemble Π_{LD} containing the single base solution from (2), then incrementally select one solution at a time from the pool to add to Π_{LD} such that the resulting ensemble has the lowest diversity. This process repeats until Π_{LD} of size 10 is obtained. This ensemble represents a "low-diversity" ensemble.
4. Construct an ensemble Π_{MD} in the same way as step 3, but incrementally select base partition that makes the resulting ensemble has medium diversity. This ensemble represents a "medium-diversity" ensemble.
5. Construct an ensemble Π_{HD} in the same way as step 3, but incrementally select a solution that makes the resulting ensemble have the highest diversity. This ensemble represents a "high-diversity" ensemble.
6. Compute the diversity of these three ensembles, i.e., $DS(\Pi_{LD})$, $DS(\Pi_{MD})$, and $DS(\Pi_{HD})$, using (17).
7. Compute the average accuracy of base partitions for these ensembles, i.e., $ACC(\Pi_{LD})$, $ACC(\Pi_{MD})$, and $ACC(\Pi_{HD})$. The average accuracy $ACC(\Pi)$ of the base partitions in Π is computed by the following equation, where Π' is the known true partition of the examined data set:

$$ACC(\Pi) = \frac{1}{10} \sum_{i=1}^{10} NMI(\pi_i, \Pi'). \qquad (18)$$

8. Apply link-based methods to these ensembles using $DC = 0.9$ and compute the accuracy of the ensemble decision, $ACC(\pi^*)$, as follows:

$$ACC(\pi^*) = NMI(\pi^*, \Pi'). \qquad (19)$$

9. Repeat steps 1 to 8 for 20 trials.

In Fig. 11, the corresponding results obtained for five real data sets are represented. To construct each plot, the results of 60 ensembles (20 for each of Π_{LD}, Π_{MD}, and Π_{HD}) are ranked in an increasing order of their diversity values. The y-axis represents NMI accuracy and the x-axis shows the ensemble indexes. This means that the first 20 indexes are the "low-diversity" ensembles. The 21st to 40th are the "medium-diversity" ensembles, and the last 20 indexes are of the "high-diversity" ensembles. The ranges of diversity are also provided in brackets for each of the diversity levels. The red line represents the average accuracy of nine link-based methods. The dashed-black line corresponds to the average accuracy of the base partitions. Base on these subfigures, the following observations are reported:

- Higher diversity values provide higher ensemble accuracies for all data sets. This result suggests that a high level of ensemble diversity is recommended for an accurate outcome.
- With the Iris, Wine, and Glass data sets, the improvement in accuracy, i.e., $ACC(\pi^*) - ACC(\Pi)$, made by the link-based approach becomes more obvious as the diversity level within the cluster ensemble increases. This especially holds for Wine, for which the link-based approach always performs better than the base clusterings. With the Ionosphere and Diabetes data sets, base clusterings outperform the link-based approach when the ensemble diversity is low to moderate. However, this is improved when the level of diversity is higher. At that point, the performance of the link-based approach exceeds that of the underlying ensembles. Again, these results confirm the preference of high ensemble diversity for link-based cluster ensembles. Please consult Section 4 of online supplemental material for further details.

5 Conclusion

This paper has presented a novel link-based approach to the cluster ensemble problem. It aims to explore and makes use of the relationships between input clusterings. This additional information that is captured as the similarity among clusters of a given ensemble allows the refined cluster-association matrix to be created. By representing these clusters as a link network, their similarity degrees can be efficiently estimated by three different link-based algorithms of WCT, WTQ, and CSM. The proposed matrix refinement is different from the previous studies that do not address the relations between ensemble members.

Two consensus methods are shown to generate the final solution from the RA matrix: Feature Based Partitioning and Bipartite Graph Partitioning. The first considers the RA matrix as a high-level data matrix to which a simple numerical partitioning techniques (e.g., k-means and PAM) can be directly applied. The second method transforms the RA matrix to a weighted bipartite graph, which is later segregated using a spectral graph partitioning technique. The empirical studies, with several ensemble settings and data sets, suggest that the link-based approach achieves superior clustering results compared to several state-of-the-art cluster ensemble techniques found in the literature. Beyond these achievements, the future work includes an extensive study regarding the behavior of other link-based similarity measures within this framework. In addition, this methodology will also be applied to specific domains such as medical and business-related data sets.

ACKNOWLEDGMENTS

The authors are grateful to the editor and anonymous reviewers for their constructive comments which have helped considerably in revising this paper. Also, the authors would like to thank X.Z. Fern and C.E. Brodley for the source code of HBGF [12], and C. Domeniconi for the implementation of LAC [8].

REFERENCES

[1] L.A. Adamic and E. Adar, "Friends and Neighbors on the Web," *Social Networks,* vol. 25, no. 3, pp. 211-230, 2003.

[2] A. Asuncion and D.J. Newman "UCI Machine Learning Repository," http://www.ics.uci.edu/~mlearn/MLRepository.html, School of Information and Computer Science, Univ. of California Irvine, 2007.

[3] H. Ayad and M. Kamel, "Finding Natural Clusters Using Multiclusterer Combiner Based on Shared Nearest Neighbors," *Proc. Int'l Workshop Multiple Classifier Systems,* pp. 166-175, 2003.

[4] T. Boongoen, Q. Shen, and C. Price, "Disclosing False Identity through Hybrid Link Analysis," *Artificial Intelligence and Law,* vol. 18, no. 1, pp. 77-102, 2010.

[5] C. Boulis and M. Ostendorf, "Combining Multiple Clustering Systems," *Proc. European Conf. Principles and Practice of Knowledge Discovery in Databases,* pp. 63-74, 2004.

[6] D. Cristofor and D. Simovici, "Finding Median Partitions Using Information-Theoretical-Based Genetic Algorithms," *J. Universal Computer Science,* vol. 8, no. 2, pp. 153-172, 2002.

[7] C. Domeniconi and M. Al-Razgan, "Weighted Cluster Ensembles: Methods and Analysis," *ACM Trans. Knowledge Discovery from Data,* vol. 2, no. 4, pp. 1-40, 2009.

[8] C. Domeniconi, D. Gunopulos, S. Ma, B. Yan, M. Al-Razgan, and D. Papadopoulos, "Locally Adaptive Metrics for Clustering High Dimensional Data," *Data Mining and Knowledge Discovery,* vol. 14, no. 1, pp. 63-97, 2007.

[9] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification,* second ed. Wiley-Interscience, Nov. 2000.

[10] S. Dudoit and J. Fridyand, "Bagging to Improve the Accuracy of a Clustering Procedure," *Bioinformatics,* vol. 19, no. 9, pp. 1090-1099, 2003.

[11] X.Z. Fern and C.E. Brodley, "Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach," *Proc. Int'l Conf. Machine Learning,* pp. 186-193, 2003.

[12] X.Z. Fern and C.E. Brodley, "Solving Cluster Ensemble Problems by Bipartite Graph Partitioning," *Proc. Int'l Conf. Machine Learning,* pp. 36-43, 2004.

[13] B. Fischer and J.M. Buhmann, "Bagging for Path-Based Clustering," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 11, pp. 1411-1415, Nov. 2003.

[14] F. Fouss, A. Pirotte, J.M. Renders, and M. Saerens, "Random-Walk Computation of Similarities between Nodes of a Graph with Application to Collaborative Recommendation," *IEEE Trans. Knowledge and Data Eng.,* vol. 19, no. 3, pp. 355-369, Mar. 2007.

[15] A.L.N. Fred, "Finding Consistent Clusters in Data Partitions," *Proc. Second Int'l Workshop Multiple Classifier Systems,* pp. 309-318, 2001.

[16] A.L.N. Fred and A.K. Jain, "Data Clustering Using Evidence Accumulation," *Proc. Int'l Conf. Pattern Recognition,* pp. 276-280, 2002.

[17] A.L.N. Fred and A.K. Jain, "Combining Multiple Clusterings Using Evidence Accumulation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 6, pp. 835-850, June 2005.

[18] L. Getoor and C.P. Diehl, "Link Mining: A Survey," *ACM SIGKDD Explorations Newsletter,* vol. 7, no. 2, pp. 3-12, 2005.

[19] A. Gionis, H. Mannila, and P. Tsaparas, "Clustering Aggregation," *Proc. Int'l Conf. Data Eng.,* pp. 341-352, 2005.

[20] S.T. Hadjitodorov, L.I. Kuncheva, and L.P. Todorova, "Moderate Diversity for Better Cluster Ensembles," *Information Fusion,* vol. 7, no. 3, pp. 264-275, 2006.

[21] D.S. Hochbaum and D.B. Shmoys, "A Best Possible Heuristic for the k-Center Problem," *Math. Operational Research,* vol. 10, no. 2, pp. 180-184, 1985.

[22] X. Hu and I. Yoo, "Cluster Ensemble and Its Applications in Gene Expression Analysis," *Proc. Asia-Pacific Bioinformatics Conf.,* pp. 297-302, 2004.

[23] N. Iam-On, T. Boongoen, and S. Garrett, "Refining Pairwise Similarity Matrix for Cluster Ensemble Problem with Cluster Relations," *Proc. 11th Int'l Conf. Discovery Science,* pp. 222-233, 2008.

[24] A.K. Jain, M.N. Murty, and P.J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys,* vol. 31, no. 3, pp. 264-323, 1999.

[25] G. Jeh and J. Widom, "SimRank: A Measure of Structural-Context Similarity," *Proc. Int'l Conf. Knowledge Discovery and Data Mining,* pp. 538-543, 2002.

[26] G. Karypis, R. Aggarwal, V. Kumar, and S. Shekhar, "Multilevel Hypergraph Partitioning: Applications in VLSI Domain," *IEEE Trans. Very Large Scale Integration Systems,* vol. 7, no. 1, pp. 69-79, Mar. 1999.

[27] G. Karypis and V. Kumar, "Multilevel k-Way Partitioning Scheme for Irregular Graphs," *J. Parallel Distributed Computing,* vol. 48, no. 1, pp. 96-129, 1998.

[28] L. Kaufman and P.J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley Publishers, 1990.

[29] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On Combining Classifiers," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 3, pp. 226-239, Mar. 1998.

[30] L.I. Kuncheva and S.T. Hadjitodorov, "Using Diversity in Cluster Ensembles," *Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics,* pp. 1214-1219, 2004.

[31] L.I. Kuncheva and D. Vetrov, "Evaluation of Stability of k-Means Cluster Ensembles with Respect to Random Initialization," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 11, pp. 1798-1808, Nov. 2006.

[32] M. Law, A. Topchy, and A.K. Jain, "Multiobjective Data Clustering," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 424-430, 2004.

[33] D. Liben-Nowell and J. Kleinberg, "The Link-Prediction Problem for Social Networks," *J. Am. Soc. Information Science and Technology,* vol. 58, no. 7, pp. 1019-1031, 2007.

[34] A. Likas, N. Vlassis, and J.J. Verbeek, "The Global k-Means Clustering Algorithm," *Pattern Recognition,* vol. 36, pp. 451-461, 2003.

[35] Z. Lin, I. King, and M.R. Lyu, "PageSim: A Novel Link-Based Similarity Measure for the World Wide Web," *Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence,* pp. 687-693, 2006.

[36] B. Minaei-Bidgoli, A. Topchy, and W. Punch, "A Comparison of Resampling Methods for Clustering Ensembles," *Proc. Int'l Conf. Machine Learning: Models, Technologies, and Applications,* pp. 939-945, 2004.

[37] E. Minkov, W.W. Cohen, and A.Y. Ng, "Contextual Search and Name Disambiguation in Email Using Graphs," *Proc. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval,* pp. 27-34, 2006.

[38] S. Monti, P. Tamayo, J.P. Mesirov, and T.R. Golub, "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data," *Machine Learning,* vol. 52, nos. 1/2, pp. 91-118, 2003.

[39] A. Ng, M. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm," *Advances in Neural Information Processing Systems,* vol. 14, pp. 849-856, 2001.
[40] N. Nguyen and R. Caruana, "Consensus Clusterings," *Proc. IEEE Int'l Conf. Data Mining,* pp. 607-612, 2007.
[41] K. Punera and J. Ghosh, "Soft Cluster Ensembles," *Proc. Advances in Fuzzy Clustering and Its Applications,* pp. 69-90, 2007.
[42] W.M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," *J. Am. Statistical Assoc.,* vol. 66, pp. 846-850, 1971.
[43] P. Reuther and B. Walter, "Survey on Test Collections and Techniques for Personal Name Matching," *Int'l J. Metadata, Semantics and Ontologies,* vol. 1, no. 2, pp. 89-99, 2006.
[44] A. Strehl and J. Ghosh, "Cluster Ensembles: A Knowledge Reuse Framework for Combining Multiple Partitions," *J. Machine Learning Research,* vol. 3, pp. 583-617, 2002.
[45] A. Struyf, M. Hubert, and P.J. Rousseeuw, "Integrating Robust Clustering Techniques in S-PLUS," *Computational Statistics and Data Analysis,* vol. 26, pp. 17-37, 1997.
[46] A.P. Topchy, A.K. Jain, and W.F. Punch, "Combining Multiple Weak Clusterings," *Proc. IEEE Int'l Conf. Data Mining,* pp. 331-338, 2003.
[47] A.P. Topchy, A.K. Jain, and W.F. Punch, "A Mixture Model for Clustering Ensembles," *Proc. SIAM Int'l Conf. Data Mining,* pp. 379-390, 2004.
[48] A.P. Topchy, A.K. Jain, and W.F. Punch, "Clustering Ensembles: Models of Consensus and Weak Partitions," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 12, pp. 1866-1881, Dec. 2005.
[49] H. Xue, S. Chen, and Q. Yang, "Discriminatively Regularized Least-Squares Classification," *Pattern Recognition,* vol. 42, no. 1, pp. 93-104, 2009.
[50] Z. Yu, H-S. Wong, and H. Wang, "Graph-Based Consensus Clustering for Class Discovery from Gene Expression Data," *Bioinformatics,* vol. 23, no. 21, pp. 2888-2896, 2007.

Natthakan Iam-On received the PhD degree in computer science from Aberystwyth University in 2011. She is a lecturer with the School of Information Technology, Mae Fah Luang University, Thailand. Her research focuses on data clustering, cluster ensembles and applications to biomedical data analysis, advance database technology, and knowledge discovery.

Tossapon Boongoen received the PhD degree in artificial intelligence from Cranfield University and worked as a postdoctoral research associate at Aberystwyth University, United Kingdom. He is a lecturer with the Department of Mathematics and Computer Science, Royal Thai Air Force Academy, Thailand. His research interests include data mining, link analysis, data clustering, fuzzy aggregation, and classification system.

Simon Garrett founded and is CEO of Aispire Consulting Ltd., having worked at Aberystwyth University in the Department of Computer Science as both a lecturer and researcher. His research has been in machine learning and clustering, which have been his interests for more than 10 years. He has been recognized for his contribution to artificial immune systems, and has done work on their ability to cluster data and find cluster centers quickly and efficiently.

Chris Price received the BSc degree in computer science from Aberystwyth University, United Kingdom, in 1979 and, after eight years building artificial intelligence systems in industry, returned to academia in 1986. He received the PhD degree in computer science from Aberystwyth University in 1994, where he was made a full professor in 1999. Much of his research has concentrated on reasoning from models to build design and diagnosis tools for use by engineers in automotive and aerospace companies.