Comparative study of matrix refinement approaches for ensemble clustering

Natthakan Iam-On · Tossapon Boongoen

Received: 25 May 2012 / Accepted: 16 March 2013 / Published online: 10 April 2013
© The Author(s) 2013

Abstract Cluster ensembles or consensus clusterings have been shown to be better than any standard clustering algorithm at improving accuracy and robustness across various sets of data. This meta-learning formalism also helps users to overcome the dilemma of selecting an appropriate technique and the parameters for that technique. Since founded, different research areas have emerged with the common purpose of enhancing the effectiveness and applicability of cluster ensembles. These include the selection of ensemble members, the imputation of missing values, and the summarization of ensemble members. In particular, this paper is set to provide the review of different matrix refinement approaches that have been recently proposed in the literature for summarizing information of multiple clusterings. With various benchmark datasets and quality measures, the comparative study of these novel techniques is carried out to provide empirical findings from which a practical guideline can be drawn.

Keywords Cluster ensemble · Multiple clusterings · Summarization · Information matrix

1 Introduction

Although cluster analysis has proven useful in many different fields; e.g. customer relationship management (Wu et al. 2005), image processing (Costa and de Andrade Netto 1999),

Editors: Emmanuel Müller, Ira Assent, Stephan Günnemann, Thomas Seidl, and Jennifer Dy.

Electronic supplementary material The online version of this article (doi:[10.1007/s10994-013-5342-y](https://doi.org/10.1007/s10994-013-5342-y)) contains supplementary material, which is available to authorized users.

N. Iam-On (✉)

School of Information Technology, Mae Fah Luang University, Chiang Rai 57100, Thailand
e-mail: nt.iamon@gmail.com

T. Boongoen

Department of Mathematics and Computer Science, Royal Thai Air Force Academy, Bangkok 10220, Thailand
e-mail: tossapon_b@rtaf.mi.th

pattern recognition (Jain et al. 2000) and recommender systems (Kim and Ahn 2008); significant challenges remain. In fact, the performance of most clustering techniques are highly data dependent such that a specific clustering model may produce an acceptable result for one set of data, but possibly become ineffective for others. Generally, there are two major problems inherent to clustering algorithms. First, different techniques discover different structures (e.g., cluster size and shape) from the same set of data objects (Duda et al. 2000; Fred and Jain 2005; Xue et al. 2009). For example, k -means which is probably the best known technique is suitable for spherical-shape clusters, while single-linkage hierarchical clustering is effective to detect connected patterns. This is due to the fact that each individual algorithm is designed to optimize a specific criterion. Second, a single clustering algorithm with different parameter settings can also reveal various structures on the same dataset. A specific setting may be good for a few, but not all datasets. A user encounters these challenges, which consequently make the selection of a proper clustering technique very difficult.

A solution to this dilemma remains an ultimate goal. In order to accomplish this, researchers invented the methodology of combining different clusterings into a single consensus clustering. This process that is widely known as ‘cluster ensembles’ can provide more robust and stable solutions across different domains and datasets (Fred and Jain 2005; Iam-On et al. 2010, 2011; Topchy et al. 2005). However, modelling a mechanism that is effective for integrating multiple data partitions in a cluster ensemble is far from trivial. This task is difficult since there is no well defined correspondence between the different clustering results. The practice of cluster ensembles is still in its early stage. Several open questions are still to be investigated and resolved. One is related to the generation (or selection) method that improves the quality of cluster ensembles. Intuitively, the performance of cluster ensembles depends on both the quality and the diversity of ensemble members (Fern and Brodley 2003). This assumption is empirically verified by Hadjitodorov et al. (2006), Kuncheva and Hadjitodorov (2004). The experimental findings suggest that more diverse ensembles provide the better quality of final clustering result, as compared to less diverse ensembles. Recent studies have also reported the improved performance of existing cluster ensemble methods through a diversity-driven determination of base clusterings’ importance (Azimi and Fern 2009; Fern and Lin 2008; Naldi et al. 2013). In addition, missing values have been recognized as a difficulty for cluster analysis and other general data mining tasks. Such a problem is under intense study, especially in the research area of microarray data analysis (see Aittokallio 2010 for a survey). A number of different techniques have been developed to impute missing values prior (Celton et al. 2010) and during (Kim et al. 2007) a clustering process.

Another topic of high interest among researchers is how to enhance the performance of consensus methods, which make an efficient use of information embedded with an ensemble and thus deriving the meaningful final data partition. Different types of matrices have been devised to summarize information of multiple base clusterings. These categories of representation dictate the technique used for generating the consensus clustering, namely direct (Fischer and Buhmann 2003), feature-based (Topchy et al. 2005), pairwise similarity based (Fred and Jain 2005) and graph-based (Strehl and Ghosh 2002) approaches, respectively. The effectiveness of initial matrices has been limited by either high complexity of matrix creation or the difficulty with categorical content. As an alternative, the matrix called ‘binary cluster association (BA)’ has been proposed by Fern and Brodley (2004), Strehl and Ghosh (2002), such that the generation of BA matrix is more efficient than the previous. Despite the reported improvement, BA matrix is generally sparse with many entries being ‘nil’ and represented as ‘0’. This may degrade the quality of final clustering, hence motivating a new research subject of refining the matrix content. Several studies have applied different matrix refinement approaches to successfully resolve the aforementioned problem. Given such

an encouragement, the paper is set to provide the review of this interesting research, including the comparative study from which a practical guideline can be formulated. Specific to the growing community interested in multiple clusterings or ‘MultiClust’, this empirical investigation provides both theoretical basis and recent development of combining multiple clustering solutions. In contrast to the recent review on cluster ensembles (Ghosh and Acharya 2011) that covers a broad scope of techniques, the present study focuses on a specific family of matrix refinement methods for ensemble clustering, with great details and an extensive comparative study. As a result, it is an effective platform to launch the application of advance approaches for new analytical problems.

The rest of this paper is organized as follows. To provide background of the subjects discussed hereafter, basic concepts and practice of cluster ensembles are presented in Sect. 2. These include different strategies used for the generation of ensembles, approaches to summarization of multiple clusterings, and a variety of consensus techniques that shape the final decision. Following that, Sect. 3 gives a thorough review on the subject of matrix refinement approaches recently proposed in the literature to improve the quality of summarization. The comparative study regarding performance of the methods previously identified on benchmark datasets and validity indices is included in Sect. 4. The paper is concluded in Sect. 5 with a perspective of future research.

2 Basis of cluster ensembles

The main objective of cluster ensembles is to combine multiple clusterings such that the resulting accuracy superior to those of individual clusterings is obtained. Studies on developing cluster ensemble methods have shown that cluster ensembles achieve the benefits beyond what standard algorithms can provide (Domeniconi and Al-Razgan 2009; Fred and Jain 2005; Gionis et al. 2007). This approach has been successfully used for many application problems, especially the analysis of cancer gene expression data (Iam-On et al. 2010; Kim et al. 2009; Monti et al. 2003; Yu et al. 2007). To set the scene for the methods discussed in later sections, this section first presents the basic concepts of cluster ensembles.

2.1 Problem formulation

Let $X = \{x_1, \dots, x_N\}$ be a set of N data points, where each $x_i \in X$ is represented by a vector of d attribute values, i.e., $x_i = \{x_1^i, \dots, x_d^i\}$. Also, let $\Pi = \{\pi_1, \dots, \pi_M\}$ be a cluster ensemble with M base clusterings, each of which is referred to as an ‘ensemble member’. Each base clustering returns a set of clusters $\pi_g = \{C_1^g, C_2^g, \dots, C_{k_g}^g\}$, such that $\bigcup_{t=1}^{k_g} C_t^g = X$, where k_g is the number of clusters in the g th clustering. For each $x_i \in X$, $C^g(x_i)$ denotes the cluster label in the g th base clustering to which data point x_i belongs, i.e., $C^g(x_i) = 't'$ (or ‘ C_t^g ’) if $x_i \in C_t^g$. The problem is to find a new partition $\pi^* = C_1^*, \dots, C_k^*$, where k denotes the number of clusters in the final clustering result, of a data set X that summarizes the information from the cluster ensemble Π .

The general framework of cluster ensembles is shown in Fig. 1. Essentially, solutions achieved from different base clusterings are aggregated to form a final partition. This meta-level method involves two major tasks of: (i) generating a cluster ensemble, and (ii) producing the final partition (normally referred to as a *consensus function*).

Fig. 1 The basic process of cluster ensembles. It first applies multiple base clusterings to a dataset X to obtain diverse clustering decisions ($\pi_1 \dots \pi_M$). Then, these solutions are combined to establish the final clustering result (π^*) using a consensus function

2.2 Ensemble generation strategies

It has been shown that ensembles are most effective when constructed from a set of predictors whose errors are dissimilar (Kittler et al. 1998). To a great extent, diversity amongst ensemble members is introduced to enhance the result of an ensemble (Kuncheva and Vetrov 2006). Specific to data clustering, the results obtained with any single algorithm over many iterations are usually very similar. Several heuristics have been proposed to introduce artificial instabilities in clustering algorithms, hence the diversity within a cluster ensemble. The following ensemble generation strategies yield different clusterings of the same data, by exploiting different cluster models and different data subsets.

- *Homogeneous ensembles*: Base clusterings are created using repeated runs of a single clustering algorithm, with several sets of input parameters. In particular, the k -means technique has often been employed with a random initialization of cluster centers (Fred and Jain 2002, 2003, 2005; Gionis et al. 2007; Iam-On et al. 2008; Topchy et al. 2004a). Other non-deterministic clustering techniques (results of multiple runs are dissimilar) such as PAM (Kaufman and Rousseeuw 1990) can also be used to form a homogeneous ensemble.
- *Different- k* : The output of clustering algorithm is dependent on the initial choice of cluster numbers k . To acquire diversity, base clusterings are created using a specific value of k or randomly selected k from a pre-specified interval. Intuitively, k should be greater than the expected number of clusters and the common rule-of-thumb is $k = \sqrt{N}$ (Fred and Jain 2005; Hadjitodorov et al. 2006; Iam-On et al. 2008; Kuncheva and Vetrov 2006).
- *Random subspace/sampling*: A cluster ensemble can also be achieved by applying different subsets of initial data to base clusterings. It is assumed that each clustering algorithm can provide different levels of performance for different subsets of a dataset (Domeniconi and Al-Razgan 2009). In practice, these can be obtained by projecting data onto different subspaces (Fern and Brodley 2003; Topchy et al. 2003), choosing different subsets of features (Strehl and Ghosh 2002; Yu et al. 2007), or using

data sampling techniques (Dudoit and Fridyand 2003; Fischer and Buhmann 2003; Minaei-Bidgoli et al. 2004).

An artificial diversity within an ensemble Π can be achieved by generating base clustering results from different perturbed variations of dataset X . To this extent, a random projection method (Bingham and Mannila 2001) is objectively used by Fern and Brodley (2003) and Topchy et al. (2003) to create such a transformed data matrix $X' \in \mathbb{R}^{N \times d'}$ from the original X , where $d' < d$. It is also possible to create different data subspaces each of which contains a randomly selected subset of original attributes (Avogadri and Valentini 2009; Yu et al. 2007). Each data subspace X' is generated by firstly defining $d' = d'_{min} + \lfloor \alpha(d'_{max} - d'_{min}) \rfloor$; where $\alpha \in [0, 1]$ is a uniform random variable, d'_{min} and d'_{max} are the lower and upper bounds of the generated subspace, respectively. Following the study of Yu et al. (2007), d'_{min} and d'_{max} are set to $0.75d$ and $0.85d$. An attribute is selected one by one from the pool of d attributes, until the collection of d' is obtained. In particular, the index of each randomly selected attribute is determined by $h = \lfloor 1 + \beta d \rfloor$, where h denotes the h^{th} attribute in the pool of d attributes and $\beta \in [0, 1)$ is a uniform random variable.

In addition to using data subspaces, an ensemble can also be created by applying a selected clustering algorithm(s) to a set of data perturbations. Specific to the strategy employed by Kim and Lee (2007) and Monti et al. (2003), each base clustering is obtained with a data subset that contains randomly selected 80 % of original data points. In the studies of Dudoit and Fridyand (2003) and Fischer and Buhmann (2003), perturbed data sets (of the same size as the original data) are obtained using the bootstrapping (or bagging) resampling scheme (Breiman 1996).

- *Heterogeneous ensembles*: As an alternative, heterogeneous ensembles may be exploited, where diversity is induced by allowing each base clustering to be generated using different clustering algorithms (Ayad and Kamel 2003; Hu and Yoo 2004; Law et al. 2004). Multiple algorithms can complement each other. Thus, combining clustering results based on multiple clustering techniques can assure better data clusterings. This approach is adapted by many ensemble algorithms, for example, that proposed by Gionis et al. (2007) applies single linkage, average linkage, complete linkage, Ward's clustering and k -means to generate the ensembles.
- *Mixed strategy*: In addition to using one of the aforementioned strategies, any combination of them can be applied as well. This can be found in Strehl and Ghosh (2002), where several clustering algorithms are used with multiple subspaces of data. Similarly, Monti et al. (2003) apply hierarchical clustering with average-linkage and the self organizing map (SOM) with different sets of sampled data, while Domeniconi and Al-Razgan (2009) generate the ensembles using their subspace clustering algorithm with different input conditions. In addition, Fred and Jain (2006) employ several strategies—applying three algorithms (k -means, single-linkage and spectral algorithm), with various initial settings, to multiple subsampled data.

2.3 Consensus functions

Having obtained the cluster ensemble, a variety of consensus functions have been developed and made available for deriving the ultimate data partition. Each consensus function utilizes a specific form of information matrix, which summarizes the base clustering results. From the cluster ensemble shown in Fig. 2(a), three general types of such ensemble-information matrix can be constructed. Firstly, the *label-assignment* matrix (e.g., see Fig. 2(b) for an example), of dimension $N \times M$, represents cluster labels that are assigned to each data point by

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

different base clusterings. Secondly, the *pairwise similarity* matrix (e.g., Fig. 2(c)), of size $N \times N$, summarizes co-occurrence statistics amongst data points. Furthermore, the *binary cluster-association (BA)* matrix, like the one shown in Fig. 2(d), provides a cluster-specific view of the original label-assignment matrix. The association degree that a data point belonging to a specific cluster is either 1 or 0. With this background, a large number of different consensus functions found in the literature can be classified to four major categorizations.

- *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 (Fig. 2(b)), which is utilized to formulate the final solution (Boulis and Ostendorf 2004; Nguyen and Caruana 2007; Topchy et al. 2005). For instance, the technique of Boulis and Ostendorf (2004) makes use of Linear Programming to find a correspondence between the labels of base clusterings and those of the optimal final-clustering. In addition, the aggregation of multiple clustering results has been considered as a maximum likelihood estimation problem, and EM algorithms (Nguyen and Caruana 2007; Topchy et al. 2004a, 2005) have been proposed for finding the consensus clustering.
- *Direct approach*: This is based on relabelling π_g and searching for the π^* that has the best match with all π_g , $g = 1 \dots M$ (Fischer and Buhmann 2003). 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. 2(b)).
- *Pairwise similarity approach*: It creates a matrix, containing the pairwise similarity among data points, to which any similarity-based clustering algorithm (e.g., hierarchical clustering) can be applied (Fred and Jain 2005). Given an ensemble Π , an $N \times N$ similarity matrix is constructed for each base clustering, denoted as S_g , $g = 1 \dots M$. Each entry in this matrix represents the relationship between two data points. If they are assigned to the same cluster, the entry will be 1, 0 otherwise. More precisely, the similarity between two data points $x_i, x_j \in X$ from the g th ensemble member can be computed as

$$S_g(x_i, x_j) = \begin{cases} 1 & \text{if } C^g(x_i) = C^g(x_j) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Following that, M similarity matrices of $S_1, \dots, S_M$ are merged to form a so-called ‘co-association (CO)’ matrix (Fred and Jain 2005)—see Fig. 2(c) for an example. Each entry

in the CO matrix represents the similarity between any two data points, which is a ratio of a number of ensemble members in which they are assigned to the same cluster to the total number of ensemble members. Formally, each entry of such a matrix $CO(x_i, x_j)$, $x_i, x_j \in X$ is defined as

$$CO(x_i, x_j) = \frac{1}{M} \sum_{g=1}^M S_g(x_i, x_j) \quad (2)$$

- *Graph-based approach*: A number of methods following this approach make use of the graph representation to solve the cluster ensemble problem (Fern and Brodley 2004; Strehl and Ghosh 2002). Specific to the consensus methods of Strehl and Ghosh (2002), a graph representing the similarity amongst data points is created from a pairwise matrix similar to that given in Fig. 2(c). To achieve the final clustering result, this graph is divided into a definite number of approximately equal-sized partitions, using a graph partitioning technique called METIS (Karypis and Kumar 1998). In addition, the binary cluster-association (BA) matrix shown in Fig. 2(d) has also been used for the generation of a bipartite graph whose vertices represent both data points and clusters. According to Fern and Brodley (2004), the solution to a cluster ensemble problem is to divide this graph using either METIS or Spectral graph partitioning (SPEC) (Ng et al. 2001).

3 Matrix refinement approaches to summarize multiple clusterings

The first wave of research on cluster ensembles are mostly motivated by the preceding work on classifier ensembles. A combination strategy such as ‘voting’, which has proven effective for classifier ensembles (Bauer and Kohavi 1999; Lam and Suen 1997), is reused with a simple label-assignment matrix being utilized as a tool for summarizing base clusterings (Fischer and Buhmann 2003). However, such practice is not directly applicable to the cluster ensemble problem where a priori class information is not available. The cluster labels in different data partitions π_g , $g = 1 \dots M$ are arbitrary. As a result, a mechanism that finds ‘label correspondence’ and re-labels each partition in accordance with a reference partition, is necessary for developing such a voting model. Most methods in this category require the number of clusters in each base partition to be k , i.e., $k_g = k$, $g = 1 \dots M$.

Along the development of an effective proximity measure for categorical data, and the corresponding clustering algorithm, another selection of cluster ensemble methods obtains the final partition by coupling such a categorical clustering with the label-assignment matrix. Despite its simplicity, the use of this matrix, either by the former ‘direct’ or the latter ‘feature-based’ approaches, can improve the overall clustering quality only marginally. Specific to the empirical study of Iam-On et al. (2011), the use of random k values in the generation of ensemble members usually results in higher accuracy, as compared to that exploitation of true k in any direct approach. In addition, the quality of a feature-based model highly depends on the categorical clustering algorithm, hence its underlying proximity measure, used to create the consensus partition. However, it has been difficult to formulate an effective proximity metric for categorical data (Iam-On et al. 2012). This leads to many attempts that transform the data domain from categorical to numerical—for instance, the ‘market-basket’ representation of nominal data (Tan et al. 2005). As a result, the pairwise-similarity matrix (i.e., ‘co-association (CO)’ matrix (Fred and Jain 2005), ‘consensus’ matrix (Monti et al. 2003), ‘similarity’ matrix (Strehl and Ghosh 2002) and ‘agreement’ matrix (Swift et al. 2004)) is created to summarize the co-occurrence statistics among all pairs of data points

(Fred and Jain 2005). This information matrix is effective to aggregate details of different data partitions, but with rather high space and time complexity— $O(N^2)$.

An alternative approach (Strehl and Ghosh 2002; Fern and Brodley 2004) to pairwise-similarity methods makes use of the binary cluster-association matrix (BA). The time complexity for creating a BA is $O(NP)$, where P denotes the number of clusters in an ensemble. This meta-level matrix, i.e., $BA \in \{0, 1\}^{N \times P}$, summarizes the cluster-data point relations occurring in the examined ensemble $\Pi = \{\pi_1, \dots, \pi_M\}$. Each entry $BA(x_i, cl) \in \{0, 1\}$ represents a *crisp* association degree that a data point x_i is in the cluster $cl \in \Pi$. Note that, for any clustering $\pi_g \in \Pi$, a data point $x_i \in X$ is a member of only one cluster. In other words,

$$\sum_{\forall cl \in \pi_g} BA(x_i, cl) = 1, \quad (3)$$

which implies the following:

$$\sum_{\forall cl \in \Pi} BA(x_i, cl) = M \quad (4)$$

It is shown in Fig. 2 that a large number of entries in the BA matrix are nil, with each being simply presented as ‘0’. Intuitively, this sparseness characteristic of BA which is commonly encountered across ensemble types may limit the quality of a data partition generated by any consensus function. This problem occurs when relations between different clusters of a base clustering are originally assumed to be nil. In the last 6–7 years, a group of cluster ensemble methods have been put forward with a common objective of refining the content of the BA matrix, i.e., transforming its value domain from crisp to continuous. Intuitively, this practice can reduce the effect of matrix sparseness, hence improving the chance of obtaining an accurate result. The followings describe four classes of matrix refinement methodology found in the literature for summarizing multiple base clusterings.

3.1 Probability based method

The initial attempt to tackle the aforementioned shortcoming is introduced by Topchy et al. (2005), namely Quadratic Mutual Information (QMI). This cluster ensemble method searches for a ‘median’ partition that is the most similar to those data partitions generated by ensemble members. This is achieved by maximizing the measure of QMI, which determines the quality of the final clustering result. In particular, QMI and CU (Category Utility) that is employed by the conceptual clustering (COBWEB) algorithm (Fisher 1987) give the same consensus clustering criterion (see proofs and further details in Topchy et al. 2005). Also, it has been demonstrated by Mirkin (2001) that the maximization of CU is equivalent to minimization of the square-error criterion of k -means if the number of clusters in target partition is fixed. In particular, the label-assignment matrix (Fig. 2(b)) acquired from the cluster ensemble under examination is firstly converted into its equivalent Binary Cluster-Association (BA) matrix (Fig. 2(d)) counterpart. It is then transformed to another numerical variation (*TMB*) to which k -means can be effectively applied. The value of each $TMB(x_i, cl)$, $\forall x_i \in X, \forall cl \in \pi_g, g = 1 \dots M$ can be defined by

$$TMB(x_i, cl) = BA(x_i, cl) - p(cl), \quad (5)$$

where $p(cl)$ is simply estimated as follows.

$$p(cl) = \frac{\sum_{j=1}^N BA(x_j, cl)}{N} \quad (6)$$

Fig. 3 TMB matrix that is generated from the ensemble shown in Fig. 2

	C_1^1	C_2^1	C_3^1	C_1^2	C_2^2
x_1	0.6	-0.4	-0.2	0.6	-0.6
x_2	0.6	-0.4	-0.2	-0.4	0.4
x_3	-0.4	0.6	-0.2	0.6	-0.6
x_4	-0.4	0.6	-0.2	-0.4	0.4
x_5	-0.4	-0.4	0.8	-0.4	0.4

π_1			π_2		
	C_1^1	C_2^1		C_1^2	C_2^2
x_1	0.6	0.4	x_1	0.7	0.3
x_2	0.7	0.3	x_2	0.2	0.8
x_3	0.1	0.9	x_3	0.6	0.4
x_4	0.0	1.0	x_4	0.1	0.9
x_5	0.4	0.6	x_5	0.0	1.0

(a)
(b)

Fig. 4 Examples of (a) cluster ensemble of fuzzy clusterings and (b) the corresponding FCM

With the example shown in Fig. 2; $p(C_1^1) = 0.4$, $p(C_2^1) = 0.4$, $p(C_3^1) = 0.2$, $p(C_1^2) = 0.4$, and $p(C_2^2) = 0.6$, respectively. As such, the resulting TMB can be shown in Fig. 3.

3.2 Fuzzy based method

In contrast to the previous framework and most of the cluster ensemble methods found in the literature, the studies of Al-Razgan et al. (2008); Punera and Ghosh (2007) are not based on crisp base clusterings. Instead, a soft or fuzzy clustering technique, such as fuzzy c -means (Bezdek 1981; Bezdek and Hathaway 1988), is exploited to create an ensemble. For any base clustering $\pi_g \in \Pi$ where $\pi_g = \{C_1^g, \dots, C_{k_g}^g\}$, each data point $x_i \in X$ is assigned with membership degrees $M_i^g = \{m_{i1}^g, \dots, m_{ik_g}^g\}$ to different clusters in π_g , where

$$\sum_{t=1}^{k_g} m_{it}^g = 1 \quad (7)$$

With this information, the creation of a soft information matrix $FCM \in [0, 1]^{N \times P}$ is straight forward. That is, the value of $FCM(x_i, C_p^g)$, $\forall C_p^g \in \pi_g$ can be simply defined by the following. To illustrate this approach, an ensemble with two fuzzy clustering members (i.e., π_1 and π_2) and the resulting FCM matrix are presented in Fig. 4(a) and (b), respectively.

$$FCM(x_i, C_p^g) = m_{iC_p^g}^g \quad (8)$$

3.3 Distance based method

Another method is developed as the by-product of a new soft-subspace clustering model (Domeniconi and Al-Razgan 2009). It generates a BA-alike information matrix, WDM , from

which the final clustering can be effectively determined. Unlike the conventional BA in which each entry is determined by the underlying label assignment, an entry $WDM(x_i, cl)$ is estimated from the distance between data point $x_i \in X$ and center of the cluster $cl \in \Pi$. For each base clustering $\pi_g \in \Pi$, the value of $WDM(x_i, cl)$, $\forall cl \in \pi_g$ can be defined by

$$WDM(x_i, cl) = \frac{D_i - d(x_i, \bar{cl}) + 1}{k_g D_i + k_g - \sum_{\forall cl' \in \pi_g} d(x_i, \bar{cl}')}, \quad (9)$$

where $d(x_i, \bar{cl})$ is the distance between data point x_i and $\bar{cl}$, that is center (or centroid) of the cluster cl . In addition, D_i can be specified as

$$D_i = \max_{\forall cl \in \pi_g} d(x_i, \bar{cl}) \quad (10)$$

According to Domeniconi and Al-Razgan (2009), the distance $d(x_i, \bar{cl})$ can be defined as follows

$$d(x_i, \bar{cl}) = \sqrt{\sum_{s=1}^d w_s^{cl} (x_s^i - \bar{cl}_s)^2} \quad (11)$$

here $w_s^{cl} \in [0, 1]$ is the weight of the s th attribute that is specific to the cluster $cl \in \pi_g$, x_s^i denotes value of the s th attribute of data x_i , and $\bar{cl}_s$ denotes the s th attribute value of the cluster center $\bar{cl}$. For any $cl \in \pi_g$,

$$\sum_{s=1}^d w_s^{cl} = 1 \quad (12)$$

The set of cluster-specific weights is systematically obtained from a so-called ‘soft sub-space clustering’ technique such as LAC (Locally Adaptive Clustering; Domeniconi et al. 2007). This method extends the conventional k -means by iteratively revising cluster-specific attribute weights that allow more compact clusters to be obtained. LAC searches for the partition $\pi = \{C_1, \dots, C_k\}$ of dataset X into k clusters that minimizes the following objective function.

$$J(U, Z, W) = \sum_{l=1}^k \sum_{s=1}^d [w_s^l O_s^l + h w_s^l \log w_s^l], \quad (13)$$

where $\sum_{l=1}^k u_{il} = 1$, $U \in \mathbb{R}^{N \times k}$ is a matrix in which each entry u_{il} represents a membership degree that data point $x_i \in X$ has with cluster $C_l \in \pi$ ($u_{il} \in \{0, 1\}$ and $u_{il} \in [0, 1]$ for crisp and soft clustering, respectively). In addition, $Z = \{z_1, \dots, z_k\}$ denotes a vectors representing the centroids of k clusters, $|C_l|$ is the cardinality of the cluster C_l , $O_s^l = \frac{1}{|C_l|} \sum_{x_i \in C_l} (x_s^i - z_s^l)^2$ and $h \geq 0$ is the constant that controls the relative differences between dimension weights. In each iteration of the k -means alike process, W is updated by

$$w_s^l = \frac{\exp(-\frac{O_s^l}{h})}{\sum_{l=1}^d \exp(-\frac{O_s^l}{h})} \quad (14)$$

3.4 Link based method

Most of the cluster ensemble methods found in the literature, including those presented above, make use of information available in an ensemble only at a coarse level (Iam-On et al. 2008). The final result is typically generated from a knowledge pool or a meta-level information matrix, which is simply created by stacking up ensemble members' results. The relations between these decisions/clustering have been unfortunately overlooked. Very few attempts, e.g., Strehl and Ghosh (2002), have been made to bring in this information, but only implicitly via a graph representation scheme. In particular, the connection between graph elements corresponding to clusters in an ensemble is restricted to a measurement by shared content. As a result, many associations, especially those within each data partition, are not addressed and used to their true potential.

Inspired by such observations, Iam-On et al. (2010) introduces a novel link-based approach to summarize base clusterings. It explicitly models base clustering results as a link network from which the relations between and within these decisions can be systematically obtained. Having acquired such information, the refined cluster-association ($RA \in [0, 1]^{N \times P}$) matrix, which is a soft variation of the original BA, is created as follows. Given a BA matrix, value of nil associations ('0') are approximated from known ones ('1'), whose association degrees are preserved within the resulting RA matrix. In other words, $\forall x_i \in X, cl \in \Pi$:

$$BA(x_i, cl) = 1 \rightarrow RA(x_i, cl) = 1 \quad (15)$$

For each clustering $\pi_g, g = 1 \dots M$ and their corresponding clusters $C_1^g, \dots, C_{k_g}^g$, the association degree $RA(x_i, cl) \in [0, 1]$ that data point $x_i \in X$ has with each cluster $cl \in \{C_1^g, \dots, C_{k_g}^g\}$ is estimated as follows:

$$RA(x_i, cl) = \begin{cases} 1 & \text{if } cl = C_*^g(x_i) \\ \text{sim}(cl, C_*^g(x_i)) & \text{otherwise} \end{cases} \quad (16)$$

where $C_*^g(x_i)$ is the cluster label (corresponding to a particular cluster of the clustering π_g) to which data point x_i has been assigned (i.e., $BA(x_i, C_*^g(x_i)) = 1$). In addition, $\text{sim}(C_x, C_y) \in [0, 1]$ denotes the similarity between any two clusters $C_x, C_y \in \pi_g$, which can be discovered using the link-based similarity algorithms presented in the following sections.

Give this definition, the following holds $\forall x_i \in X, \pi_g \in \Pi$:

$$1 \leq \sum_{\forall cl \in \pi_g} RA(x_i, cl) \leq k_g \quad (17)$$

Unlike the measure of fuzzy membership or conventional probability, the typical constraint of $\sum_{\forall cl \in \pi_g} RA(x_i, cl) = 1$ is not applied for re-scaling association values within the RA. According to Iam-On et al. (2011), this local normalization will significantly distort the true semantics of known associations ('1'), such that their magnitudes become dissimilar, different from one clustering to another.

3.4.1 Graph based representation of a cluster ensemble

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

Fig. 5 Example of cluster network, where $V = \{C_1^1, C_2^1, C_3^1, C_1^2, C_2^2\}$ and each edge is marked with its weight

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:

$$|w_{xy}| = \frac{|L_x \cap L_y|}{|L_x \cup L_y|}, \quad (18)$$

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 $|w_{xy}|$ is equivalent to $|w_{yx}|$, $\forall v_x, v_y \in V$.

Figure 5 shows the network of clusters that is generated from the example given in Fig. 2. Note that circle nodes represent clusters and edges exist only when the corresponding weights are non-zero. This graph represents the similarity amongst 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., that between clusters C_1^1 and C_2^1 , are simply nil.

It is noteworthy that a similar graph that represents the equivalent information to a BA matrix has also been employed by the Meta-Clustering Algorithm (MCLA; Strehl and Ghosh 2002). However, that graph is only used to represent the content-based similarity amongst 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 (Karypis and Kumar 1998) to partition this graph into meta-clusters from which the final data partition is approximated. In contrary, 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. Based on this representation, the similarity between clusters C_x and C_y can be determined as that between vertices $v_x, v_y \in V$. The following link-based similarity algorithms, namely Weighted Connected-Triple (WCT) and Weighted Triple-Quality (WTQ), are particularly introduced for this similarity estimation.

3.4.2 Weighted connected-triple (WCT) algorithm

Shared neighbors have been widely recognized as the basic evidence to justify the similarity amongst vertices in a link network (Getoor and Diehl 2005; Liben-Nowell and Kleinberg 2007). Formally, a vertex $v_i \in V$ is a common neighbor of vertices $v_j, v_{j'} \in V$, provided that $w_{ij}, w_{ij'} \in W$. Many advanced path-based methods extend this node-based basis by taking into account common neighbors that may be many edges away from the two under examination: for instance, Connected-Path (Boongoen et al. 2010), SimRank (Jeh and Widom 2002), PageSim (Lin et al. 2006) and a variation of random walk algorithms (Fouss et al. 2007; Minkov et al. 2006). Despite reported effectiveness, these techniques are computationally expensive, or even impractical for a large data set. Henceforth, the Weighted Connected-Triple (WCT) and Weighted Triple-Quality (WTQ) algorithms are proposed by Iam-On et al. (2011) for efficiently approximating the similarity between clusters in the aforementioned link network.

Firstly, WCT extends the Connected-Triple method (Reuther and Walter 2006) that has been developed to identify ambiguous author names within a publication database, i.e., Digital Bibliography and Library Project (DBLP). This approach is built on a social network which can be 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 co-authorship relation (i.e., appearing as co-authors in one or more publications) between two involving authors. 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, denoted by $Triple = (V_{Triple}, E_{Triple})$, is a subgraph of G containing three vertices $V_{Triple} = \{v_x, v_y, v_z\} \subset V$ and two edges $E_{Triple} = \{e_{xz}, e_{yz}\} \subset E$, 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 points. According to this observation, the WCT algorithm is established. With a weighted graph $G = (V, W)$ presented in Fig. 5, 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

$$WCT_{xy}^z = \min(|w_{xz}|, |w_{yz}|) \quad (19)$$

The summation of all triples $(1 \dots q)$ between vertices v_x and v_y can be calculated by the following:

$$WCT_{xy} = \sum_{z=1}^q WCT_{xy}^z \quad (20)$$

Given these definitions, a pseudocode for the WCT measure is described in Algorithm 1. The similarity $sim(v_x, v_y)$ between clusters v_x and v_y can be estimated by

$$sim(v_x, v_y) = \frac{WCT_{xy}}{WCT_{max}} \times DC, \quad (21)$$

where $DC \in [0, 1]$ is a constant decay factor (i.e., confidence level of accepting two non-identical clusters as being similar) and WCT_{max} is defined as

$$WCT_{max} = \max_{\forall v_x, v_y \in V} WCT_{xy} \quad (22)$$

With this link-based similarity metric, i.e., $sim \in [0, 1]^{P \times P}$, $sim(v_x, v_y) \in [0, 1]$ with $sim(v_x, v_x) = 1$, $\forall v_x, v_y \in V$. It is also reflexive such that $sim(v_x, v_y)$ is equivalent to $sim(v_y, v_x)$. Following the example shown in Figs. 2, 5 and Eq. (16), the resulting RA matrix and those relations discovered by the WCT algorithm are presented in Fig. 6.

3.4.3 Weighted triple-quality (WTQ) algorithm

Unlike the WCT method that concentrates solely on the magnitude of common triples, the Weighted Triple-Quality (WTQ) algorithm aims to differentiate their significance and hence their contributions towards the underlying similarity measure. WTQ is based on the initial measure of Adamic and Adar (2003), 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 $s(p_a, p_b)$ as follows:

$$s(p_a, p_b) = \sum_{\forall z_c \in Z} \frac{1}{\log(\text{frequency}(z_c))}, \quad (23)$$

Algorithm 1 Pseudocode for the WCT measure**Input:**

$G = (V, W)$, a weighted graph
 $v_x, v_y \in V$, the two clusters under examination

Variable:

WCT_{xy} , the WCT measure of v_x and v_y
 $N_x \subset V$, the set of adjacent neighbours of v_x
 $N_y \subset V$, the set of adjacent neighbours of v_y
 $N_{xy} \subset V$, the set of neighbours shared by v_x and v_y

- (1) Set $WCT_{xy} = 0$
- (2) Find N_x such that $\forall v_a \in N_x, |w_{xa}| > 0$
- (3) Find N_y such that $\forall v_b \in N_y, |w_{yb}| > 0$
- (4) Find $N_{xy} = N_x \cap N_y$
- (5) For each $v_z \in N_{xy}$
 Compute $WCT_{xy} = WCT_{xy} + \min(|w_{xz}|, |w_{yz}|)$

Output: WCT_{xy}

Fig. 6 An example of discovered relations (using the WCT algorithm) and the resulting RA matrix, where $DC = 0.9$

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 appearing 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, Eq. (23) 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 $G = (V, W)$ presented in Fig. 5, 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:

$$WTQ_{xy}^z = \frac{1}{W_z}, \quad (24)$$

Algorithm 2 Pseudocode for the WTQ measure**Input:**

$G = (V, W)$, a weighted graph
 $v_x, v_y \in V$, the two clusters under examination

Variable:

WTQ_{xy} , the WTQ measure of v_x and v_y
 $N_x \subset V$, the set of adjacent neighbours of v_x
 $N_y \subset V$, the set of adjacent neighbours of v_y
 $N_{xy} \subset V$, the set of neighbours shared by v_x and v_y
 $N_z \subset V$, the set of adjacent neighbours of $v_z \in N_{xy}$

- (1) Set $WTQ_{xy} = 0$
- (2) Find N_x such that $\forall v_a \in N_x, |w_{xa}| > 0$
- (3) Find N_y such that $\forall v_b \in N_y, |w_{yb}| > 0$
- (4) Find $N_{xy} = N_x \cap N_y$
- (5) For each $v_z \in N_{xy}$
 - (5.1) Find N_z such that $\forall v_c \in N_z, |w_{zc}| > 0$
 - (5.2) Compute $WTQ_{xy} = WTQ_{xy} + \frac{1}{\sum_{v_c \in N_z} |w_{zc}|}$

Output: WTQ_{xy}

where

$$W_z = \sum_{\forall v_t \in N_z} |w_{zt}| \quad (25)$$

Here $N_z \subset V$ denotes the set of vertices that is directly linked to the vertex v_z , such that $\forall v_t \in N_z, w_{zt} \in W$. The accumulative WTQ score from all triples $(1 \dots q)$ between vertices v_x and v_y can be approximated by the following. Also, a pseudocode for the WTQ measure is described in Algorithm 2.

$$WTQ_{xy} = \sum_{z=1}^q WTQ_{xy}^z \quad (26)$$

Following that, the similarity $sim(v_x, v_y)$ between vertices v_x and v_y can be estimated by

$$sim(v_x, v_y) = \frac{WTQ_{xy}}{WTQ_{max}} \times DC, \quad (27)$$

where WTQ_{max} is the maximum WTQ_{pq} value of any two vertices $v_p, v_q \in V$. This similarity metric has the properties similar to that obtained with the WCT technique. By following the example shown in Figs. 2, 5 and Eq. (16), the resulting RA matrix and those relations discovered by the WTQ algorithm are presented in Fig. 7.

3.5 Discussion

The aforementioned approaches commonly strive for a refined variation of the original BA matrix, which is thought to be more informative for the following step of creating a con-

Fig. 7 An example of discovered relations (using the WTQ algorithm) and the resulting RA matrix, where $DC = 0.9$

Fig. 8 The scope of ensemble information used by different methods to refine BA matrix

sensus partition. To achieve that, different techniques and disciplines such as fuzzy clustering and graph theory, have been integrated into the framework of ensemble clusterings. In addition, distinct contexts of information reuse lead to another interesting comparison. Figure 8 summarizes the scope of information used by different approaches. To start with the probability-based method, only cluster-level information is exploited for the determination of a TMB matrix. A wider scope, i.e., clustering-specific details, is brought into consideration by both distance-based and fuzzy-based counterparts. With the link-based methods, this is further extended to the use of all clusterings via the graph representation.

Following that, it is also necessary to observe the differences in term of complexity, i.e. scalability to a large set of data. As such, the time complexity to create a BA matrix is $O(M)$. This is efficient since M clusterings are simply stacked up together to form the desired information matrix. See Table 1 for the complexity of other matrices.

Given these estimates, both link-based matrices are more expensive to build, especially when the underlying ensemble possesses a large number of clusters. Hence, for a time-sensitive application, other matrices like FCM and WDM should be more practical. Note that the actual running time to produce each matrix for a collection of examined datasets will be discussed later in Sect. 4. In terms of algorithmic and complexity requirements, Fig. 9 illustrates the core characteristics of the four matrix refinement approaches, which can be a useful guideline for general users. In particular, the probability-based method appears to be the most user-friendly, while the others need parameter settings, which are usually data dependant. According to the original works on LAC technique (Domeniconi et al. 2007),

Table 1 Time complexity of generating different matrices. Note that P represents the total number of clusters in an ensemble of size M , and l denotes the average number of neighbors of any vertex in a graph under examination

Matrix Type	Complexity
CO	$O(N^2)$
BA	$O(M)$
FCM	$O(M)$
TMB	$O(M + NP)$
WDM	$O(M + NP)$
RA _{WCT}	$O(M + NP + P^2l)$
RA _{WTQ}	$O(M + NP + P^2l^2)$

	Algorithms for base clusterings	Parameters	Complexity
Probability-based	Crisp clustering	n/a	Medium
Fuzzy-based	Fuzzy clustering	Fuzzy index $\delta \in (1, \infty)$	Low
Distance-based	Subspace clustering (e.g., LAC)	Weighting factor $h \in [0, \infty)$	Medium
Link-based	Crisp clustering	Decay constant DC $[0, 1]$	High

Fig. 9 Conceptual comparison of matrix refinement approaches

fuzzy c-means clustering (Bezdek 1981) and link-based cluster ensembles (Iam-On et al. 2010); $h = 9$, $\delta = 2$ and $DC = 0.9$ have been shown to be effective for several datasets.

4 Comparative study

Having explored the advanced methods for the summarization of multiple clusterings, it is interesting to observe and compare their performance on a variety of datasets, experimental settings and different quality indices. The findings from this empirical study can provide insightful information regarding the selection of a summarization method that is appropriate for a given problem.

4.1 Experimental design

To obtain a rigorous comprehension towards the effectiveness of different approaches, this section presents the methodology which is systematically designed and employed for the performance evaluation.

Table 2 Description of datasets: number of data points (N), number of attributes (d), number of classes (k) and source

Dataset	N	d	k	Source
<i>Synthetic Dataset</i>				
Difficult Doughnut	100	12	2	Kuncheva and Vetrov (2006)
2-doughnut	100	3	2	de Castro (2001)
2-banana	200	2	2	Iam-On et al. (2008)
5-gaussian	600	2	5	modified from Kuncheva and Vetrov (2006)
Fence	700	2	5	Iam-On et al. (2011)
<i>Real Dataset</i>				
Iris	150	4	3	UCI, Asuncion and Newman (2007)
Wine	178	13	3	UCI, Asuncion and Newman (2007)
E. coli	336	7	8	UCI, Asuncion and Newman (2007)
Ionosphere	351	34	2	UCI, Asuncion and Newman (2007)
Breast Cancer	683	9	2	UCI, Asuncion and Newman (2007)
Optical Digits	5,620	64	10	UCI, Asuncion and Newman (2007)
Pen Digits	10,992	16	10	UCI, Asuncion and Newman (2007)
Magic Gamma	19,020	10	2	UCI, Asuncion and Newman (2007)

4.1.1 Investigated data

A collection of thirteen datasets are included in the experiments. Table 2 summarizes the details of these datasets that are grouped into synthetic and real categories. Five synthetic datasets included in the experiments: Difficult Doughnut, 2-doughnut, 2-banana, 5-gaussian and Fence, are shown in Fig. 10(a) to 10(e), respectively. Note that the first synthetic dataset acquired from Kuncheva and Vetrov (2006) is initially created in two dimensions and later added with ten more dimensions of noise. In addition to the synthetic data collection, eight real datasets obtained from the benchmark UCI repository (Asuncion and Newman 2007) are also included.

Iris is very commonly used in cluster ensembles and other areas of data mining (Fred and Jain 2005; Topchy et al. 2004b). There are four attributes of measurements taken for each of 150 *Iris* flowers, which are the length and the width of sepal and petal, in centimeters. It consists of three classes of *Iris* plants—Setosa, Versicolour and Virginica with the class distribution of 50-50-50.

Wine contains the results of a chemical analysis of wines, which are explained by 13 attributes. This dataset includes 178 data instances of three types of wine (59, 71, and 48 entries for each class).

E. coli is a collection of 336 protein sequences of *E. coli* from SWISS-PROT, i.e., a curated protein sequence database. Based on 7 features of amino acid sequence analysis, they are classified into 8 classes of protein localization sites: inner membrane lipoproteins (imL), outer membrane lipoproteins (omL), inner membrane proteins with cleavable signal sequence (imS), other outer membrane proteins (om), periplasmic proteins (pp), inner membrane proteins with an uncleavable signal sequence (imU), inner membrane proteins without a signal sequence (im), and cytoplasmic proteins (cp).

Ionosphere is a collection of 351 radar signals, which is collected from a phased array of 16 high-frequency antennas. Free electrons in the ionosphere were targeted. These data

Fig. 10 Synthetic datasets: (a) Difficult Doughnut, (b) 2-doughnut, (c) 2-banana, (d) 5-gaussian and (e) Fence

are labelled as ‘good’ (radar returns that show some structure in the ionosphere) and ‘bad’ (those that do not). It contains 34 continuous attributes representing the time measurement of pulse and pulse number.

Breast Cancer was originally obtained from the University of Wisconsin Hospitals (Mangasarian and Wolberg 1990). The dataset consists of 699 samples described by 9 attributes of cytological characteristics. These nine features are used to categorise samples into two classes of ‘benign’ (458 samples) and ‘malignant’ (241 samples). Note that 16 samples with missing values are excluded, thus the resulting dataset used in this evaluation contains 683 instances.

Optical Digits is extracted as normalized bitmaps of handwritten digits from a preprinted form. Specific to each digit, a matrix of 8×8 is created with each element being an integer in the range of 0 to 16. The ten classes of this dataset, i.e., 0 to 9, have roughly the same amount of data points.

Pen Digits is originally collected from 250 samples produced by 44 writers. It consists of 10,992 data points, each of which is represented by 16 attributes. Like the previous dataset, the sizes of 10 classes are almost equal.

Magic Gamma has been generated to simulate registration of high energy gamma particles in a ground-based atmospheric Cherenkov gamma telescope. It contains 19,020 samples that are described by 10 different attributes. These can be categorized into two classes of ‘g’ (gamma/signal) and ‘h’ (hadron/background), with the sizes of 12,332 and 6,688, respectively.

4.1.2 Experimental setting

This set of experiments set out to investigate the quality of data partitions produced from the BA matrix and its soft variations, using different consensus functions. There are seven types of matrix to be examined. This includes the conventional pairwise similarity matrix, i.e., CO, which provides an additional baseline for those advanced alternatives.

– CO: the original co-association matrix,

- *BA*: the original matrix,
- *TMB*: the soft variation of *BA*, created by the probability based method,
- *FCM*: the soft variation of *BA*, created by the fuzzy based method,
- *WDM*: the soft variation of *BA*, created by the distance based method,
- *RA_{WCT}*: the soft variation of *BA*, created by the link based method and WCT similarity measure, and
- *RA_{WTQ}*: the soft variation of *BA*, created by the link based method and WTQ similarity measure.

The quality of the aforementioned matrices are judged through the quality of final clusterings, which are achieved by coupling each of these matrices with the following consensus functions:

- Three hierarchical methods (Han and Kamber 2000): single-linkage (SL), complete-linkage (CL) and average-linkage (AL), and
- Three partitioning methods: k-means (KM; McQueen 1967), Partitioning Around Medoids (PAM; Kaufman and Rousseeuw 1990), Spectral clustering (SPC; Ng et al. 2001), and
- Two graph partitioning methods: METIS (Karypis and Kumar 1998) and Spectral graph partitioning (gSPEC; Ng et al. 2001).

In addition, details of other experimental settings employed in this evaluation are exhibited below.

- The resulting models are investigated using two ensemble types of: Fixed- k and Random- k . Following the study of Iam-On et al. (2011), the Fixed- k scheme refers to setting of the number of clusters for each base clustering to be a predefined number $k = \sqrt{N}$. For the case of Random- k , the number of clusters is randomly chosen from the range of $\{2, \sqrt{N}\}$. In case of a large dataset with $k > 50$, it is intuitively set to 50 as to generate a meaningful partition (Iam-On et al. 2012). Note that the classical k -means is used with random initialization to create base clusterings.
- Ensemble size (M) of 10 base clusterings is experimented.
- For comparison, as in Fern and Brodley (2004), Fred and Jain (2005), Gionis et al. (2007), each method divides data points into a partition of k (the number of *true classes* for each dataset) clusters, which is then evaluated against the corresponding true partition. Note that, true classes are known for all datasets but are not explicitly used by the cluster ensemble process. They are only used to evaluate the quality of the clustering results.
- The quality of each cluster ensemble method with respect to a specific ensemble setting is generalized as the average of 50 runs. Based on the central limit theorem (CLT), the observed statistics in a controlled experiment can be justified to the normal distribution (Tijms 2004). Hence, the average of obtained results is used as the simple and effective approximation of performance.
- The constant decay factor (DC) of 0.9 is exploited with the link-based similarity algorithms, i.e., WCT and WTQ. In addition, the parameter h of LAC clustering algorithm that used to created the WDM matrix, is randomly selected from $\{1, 2, \dots, 9\}$. Also, the fuzzy index δ for the generation of FCM is set to 2. See more details in Sect. 3.
- The quality of final clusterings is evaluated using both internal and external validity indices. Firstly, the goodness of a consensus partition is estimated using Dunn (Dunn 1974) and Silhouette (Kaufman and Rousseeuw 1990) measures. These justify the quality of based purely on the characteristics of data clusters such as compactness and separability, without referring to external information (e.g., ground truth). On the other hand, provided

that the external class labels are available for all experimented datasets, the results of final clustering can be evaluated using the validity indices of Jaccard (Jain and Dubes 1998) and Normalized Mutual Information (NMI; Strehl and Ghosh 2002).

4.2 Experimental results

Based on the NMI measures, Figs. 11, 12, 13 presents the experimental results with respect to different consensus techniques of SL, CL, AL, KM, PAM, SPC, METIS and gSPEC, respectively. See the full results in Supplementary (Tables 1–8). According to the statistics shown in these figures, RA_{WCT} and RA_{WTQ} often provide the clusterings of higher quality than the base line (i.e., BA) and other soft variations. Another important observation is that the coupling of TMB matrix and SL seems to be effective for synthetic data. According to the statistics presented in these tables, the quality of both link-based matrices are generally higher than others, with WDM being the most accurate among the rest. Specific to the CO matrix, its couplings with SL and AL appear to be more effective in several datasets, as compared to the other consensus functions examined herein. In addition, similar observations have been obtained using the validity index of Jaccard. See related results in Supplementary (Tables 9–16).

It is noteworthy that both NMI and Jaccard compare the partition obtained from the consensus clustering with the collection of classes identified by domain experts, i.e., ground truth. Thus, it is also compulsory to justify the quality of ensemble models on the properties, which are preferred for good data clusters—e.g., each cluster is compact whilst clearly separable from the others. To this end, two so-called internal validity indices of Dunn and Silhouette are employed to provide additional performance-oriented statistics (see Tables 17–24 and Tables 25–32 in Supplementary for results regarding Dunn and Silhouette measures, respectively). Specific to these, FCM and WDM have shown to be as effective as the two link-based matrices. In some datasets like ‘Pen Digits’, the use of WDM leads to the best quality measures. Like before, applying SL and AL as a consensus function to CO matrix gives competitive results.

In order to compare the performance of different clustering models, the overall quality measure for a specific experiment setting (i.e., dataset and ensemble type) is obtained as the average of validity measures across 50 trials. These method-specific means may be used for the comparison purpose only to a certain extent. To achieve a more reliable assessment, the number of times (or frequencies) that one technique is ‘significantly better’ and ‘significantly worse’ (of 95 % confidence level) than the others are considered here. This comparison method has been successfully exploited by Iam-On et al. (2010) and Kuncheva (2006) to discover trustworthy conclusions from the results generated by different cluster ensemble approaches. Formally, let $\mu(i, j, \theta)$ be the average of measures of $\Phi \in \{NMI, Jaccard, Dunn, Silhouette\}$ across n runs ($n = 50$ in this evaluation) for a clustering model $i \in CM$ (CM is a set of experimented models), on a specific dataset $j \in DT$ (DT is a set of investigated datasets) and an ensemble type $\theta \in \Theta$ (Θ is a set of different ensemble types or generation strategies). In other words, $\mu(i, j, \theta)$ is estimated by

$$\mu(i, j, \theta) = \frac{1}{n} \sum_{t=1}^n \Phi_t(i, j, \theta), \quad (28)$$

where $\Phi_t(i, j, \theta)$ denotes the Φ measure obtained from the t th run of i method, in the setting of dataset j and ensemble type θ .

Fig. 11 NMI measures of examined matrices with different consensus functions, on five synthetic datasets

Fig. 12 NMI measures of examined matrices with different consensus functions, on Iris, Wine, E. coli and Ionosphere datasets

The comparison of average values (or means) to discriminate the effectiveness of examined methods may be misleading as the difference between means can be statistically insignificant at times. Thus, such an evaluation decision can be more robust using the 95 % confidence interval for the mean $\mu(i, j, \theta)$, which is defined as follows:

$$\left[\mu(i, j, \theta) - 1.96 \frac{S(i, j, \theta)}{\sqrt{n}}, \mu(i, j, \theta) + 1.96 \frac{S(i, j, \theta)}{\sqrt{n}} \right], \quad (29)$$

where $S(i, j, \theta)$ denotes the standard deviation of the Φ measures across n runs for a clustering method i over a dataset j and ensemble type θ . The statistical significance of the

Fig. 13 NMI measures of examined matrices with different consensus functions, on Breast Cancer, Optical Digits, Pen Digits and Magic Gamma datasets

difference between any two clustering models $i, i' \in CM$ over any dataset $j \in DT$ and ensemble type $\theta \in \Theta$ is found if there is no intersection between their confidence intervals of $\mu(i, j, \theta)$ and $\mu(i', j, \theta)$. For any dataset j and ensemble type θ , a clustering model i is significantly *better* than another model i' when

$$\left(\mu(i, j, \theta) - 1.96 \frac{S(i, j, \theta)}{\sqrt{n}} \right) > \left(\mu(i', j, \theta) + 1.96 \frac{S(i', j, \theta)}{\sqrt{n}} \right) \quad (30)$$

Following that, the number of times (or frequencies) that one method $i \in CM$ is significantly *better* than its competitors across all experimented datasets and ensemble types, i.e., $B(i)$,

can be estimated as follows:

$$B(i) = \sum_{\forall j \in DT} \sum_{\forall \theta \in \Theta} \sum_{\forall i' \in CM, i' \neq i} better_{\theta}^j(i, i'), \quad (31)$$

where

$$better_{\theta}^j(i, i') = \begin{cases} 1 & \text{if } (\mu(i, j, \theta) - 1.96 \frac{S(i, j, \theta)}{\sqrt{n}}) > (\mu(i', j, \theta) + 1.96 \frac{S(i', j, \theta)}{\sqrt{n}}) \\ 0 & \text{otherwise} \end{cases} \quad (32)$$

Similarly, the number of times that one model $i \in CM$ is significantly *worse* than its competitors, i.e., $W(i)$, can be computed by

$$W(i) = \sum_{\forall j \in DT} \sum_{\forall \theta \in \Theta} \sum_{\forall i' \in CM, i' \neq i} worse_{\theta}^j(i, i') \quad (33)$$

where

$$worse_{\theta}^j(i, i') = \begin{cases} 1 & \text{if } (\mu(i, j, \theta) + 1.96 \frac{S(i, j, \theta)}{\sqrt{n}}) < (\mu(i', j, \theta) - 1.96 \frac{S(i', j, \theta)}{\sqrt{n}}) \\ 0 & \text{otherwise} \end{cases} \quad (34)$$

Given these definitions, it is useful to evaluate the quality of many clustering techniques based on the frequencies of better (B) and worse (W) performance than competitors. In particular to NMI, Table 3 shows the details of better and worse statistics of examined information matrices (see Tables 33–35 in Supplementary, for similar comparisons with the other validity measures). The presented frequencies are achieved as the summarization across all datasets and ensemble types. These results consolidate the aforementioned observations, where the two link-based methods are the most accurate and WDM appears to the most effective amongst the rest. In addition, the original BA appears to be better than TMB and FCM, which suggests that they are more data dependant or less robust as compared to the others. Based on the graphical representation of B and W statistics in Fig. 14, the aforementioned trend can be clearly seen with NMI and Jaccard measures, while the performance of FCM is slightly better with respect to internal indices. Despite its high complexity, CO matrix is another effective alternative, with the quality measures being competitive to those of BA.

The algorithmic parameter that may determine the quality of data partition generated by a cluster ensemble technique, is the ensemble size (M). It is commonly perceived that the larger an ensemble is, the better the performance becomes. As shown in Fig. 15, this heuristic is usually applicable to various combinations of matrices and consensus functions, where the average NMI measures across all ensemble types and investigated datasets gradually incline to the increasing value of $M \in \{10, 20, \dots, 100\}$. This can be similarly observed using the other quality measures included in the comparative study. Apart from the performance-led comparison, it is also important to observe the differences among the investigated models in term of scalability. Table 4 presents the average computation time (in seconds), across 50 runs, for constructing each matrix from an ensemble of ten base clusterings. This is conducted on a work station (Intel Core i7 3.70 GHZ, 8 GB RAM). Note that the fixed- k scheme is used here to deliver a fair comparison on different ensembles. According to these statistics, BA and FCM are the most efficient, while both link-based methods (RA_{WCT} and RA_{WTO}) are the most expensive among the matrices under examination.

Table 3 The statistics of better (B) and worse (W) performance of each information matrix according to NMI index, categorized with respect to the consensus function. These are summarized across datasets and ensemble types

Con. Fn.	Better							Worse						
	RA _{WCT}	RA _{WTQ}	BA	FCM	WDM	TMB	CO	RA _{WCT}	RA _{WTQ}	BA	FCM	WDM	TMB	CO
SL	55	44	36	10	19	34	34	3	5	12	102	83	22	5
CL	112	101	23	17	85	24	19	8	16	74	100	41	70	72
AL	72	72	37	7	51	35	35	10	19	36	134	62	28	20
KM	87	74	46	32	39	24	50	14	26	44	92	82	71	23
PAM	97	100	23	48	67	33	28	18	19	80	84	60	70	65
SPC	75	71	46	22	50	42	33	21	22	43	109	66	34	44
METIS	72	59	37	29	32	26	41	14	22	27	74	84	47	28
gSpec	104	83	38	34	63	35	25	13	24	58	94	53	78	62
Total	674	604	286	199	406	253	265	101	153	374	789	531	420	319

Fig. 14 The better (B) and worse (W) statistics of different matrices with respect to four validity indices, summarized across all datasets, consensus functions and ensemble types

In practice, the use of cluster ensembles have been criticized for time complexity, which generally comprises three parts—the generation of ensemble, the matrix formation, and the consensus clustering, respectively. Specific to this comparative study, the complexity of the first phase is $O(NM)$, where that of each base clustering is $O(N)$. As for the matrix creation, this depends on the matrix type, which the range from $O(M)$ of BA to

Fig. 15 The relations between $M \in \{10, 20, \dots, 100\}$ and the performance of examined methods (the averages of NMI across ensemble types for each dataset), whose values are presented in x -axis and y -axis, respectively

Table 4 Computational time (seconds) required to produce dataset-specific matrices, from an ensemble of size 10

Dataset	<i>BA</i>	<i>TMB</i>	<i>FCM</i>	<i>WDM</i>	<i>RA_{WCT}</i>	<i>RA_{WTQ}</i>
Difficult Doughnut	0.0001	0.0033	0.0001	0.0035	0.0156	0.0167
2-doughnut	0.0001	0.0031	0.0001	0.0033	0.0140	0.0156
2-banana	0.0003	0.0050	0.0003	0.0052	0.0390	0.0421
5-gaussian	0.0016	0.0092	0.0017	0.0095	0.1825	0.1903
Fence	0.0020	0.0113	0.0020	0.0117	0.2424	0.2512
Iris	0.0002	0.0036	0.0002	0.0041	0.0265	0.0288
Wine	0.0003	0.0048	0.0003	0.0052	0.0359	0.0373
E. coli	0.0008	0.0067	0.0008	0.0069	0.0796	0.0827
Ionosphere	0.0008	0.0068	0.0009	0.0071	0.0827	0.0837
Breast Cancer	0.0018	0.0102	0.0018	0.0106	0.2340	0.2387
Optical Digits	0.0156	0.0412	0.0159	0.0420	4.0262	4.1964
Pen Digits	0.0390	0.0936	0.0392	0.0989	7.6937	7.9365
Magic Gamma	0.0820	0.2340	0.0831	0.2355	13.1041	13.4317

Fig. 16 Decision-support contexts for users' selection of an ensemble matrix

$O(M + NP + P^2l^2)$ of RA_{WTQ} (see related discussion in Sect. 3). Having obtained the preferred matrix, a number of consensus functions are available, each with distinct time requirements. Both partitioning techniques of KM and PAM are the most efficient with the complexity around $O(N)$. In addition, those of SPC, METIS and gSPEC are $O(N^2)$, while hierarchical clustering methods (i.e., SL, CL and AL) are the most expensive with time complexity of $O(N^3)$. Based on these, cluster ensembles may not always be suitable for all data analytical tasks, especially when allowed running time is limited. Despite this constraint, the ensemble methodology has proven effective with the quality beyond that of any standard technique. To provide a selection guideline for general users/scientists, Fig. 16 illustrates different problem contexts and the suitable matrices to summarizing multiple clusterings.

5 Conclusion

This paper has presented the review of matrix-refinement approaches to summarization of multiple clusterings. These new methods commonly aim to transform the domain of origi-

nal data matrix from crisp to continuous, such that the problem with sparseness can be eased. They achieve this goal via different practices: encoding cluster-specific probability, utilizing fuzzy base clusterings, making use of the distance to discovered centroids, and relations across clusterings. The comparative study illustrates their effectiveness and practical limitations as a general guideline.

Despite the promising findings, there are several challenges remain. The obvious question arises with the possibility of applying the link-based intuition to other matrix-refinement methods. Bringing the associations amongst multiple clusterings into consideration may improve the quality of those matrices, hence the resulting data partition. Another nontrivial work regards a better induction of diversity within cluster ensembles. It is interesting to observe the behavior of these matrices to new ensemble generation strategies, for instance, the random forest method for clustering (Shi and Horvath 2006; Shi et al. 2005), which may impose a higher diversity amongst base clusterings. In addition to the homogeneous ensemble of k -means which is currently used for its efficiency and quality reported in the literature, the investigation of heterogeneous ensembles (i.e., different algorithms as base clusterings) may project a number of useful issues and possible improvements.

References

- Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. *Social Networks*, 25(3), 211–230.
- Aittokallio, T. (2010). Dealing with missing values in large-scale studies: microarray data imputation and beyond. *Briefings in Bioinformatics*, 11(2), 253–264.
- Al-Razgan, M., Domeniconi, C., & Barbara, D. (2008). Random subspace ensembles for clustering categorical data. In *Supervised and unsupervised ensemble methods and their applications* (pp. 31–48). Berlin: Springer.
- Asuncion, A., & Newman, D. J. (2007). UCI machine learning repository. <http://www.ics.uci.edu/~mllearn/MLRepository.html>.
- Avogadri, R., & Valentini, G. (2009). Fuzzy ensemble clustering based on random projections for DNA microarray data analysis. *Artificial Intelligence in Medicine*, 45, 173–183.
- Ayad, H., & Kamel, M. (2003). Finding natural clusters using multicluseter combiner based on shared nearest neighbors. In *Proceedings of international workshop on multiple classifier systems* (pp. 166–175).
- Azimi, J., & Fern, X. (2009). Adaptive cluster ensemble selection. In *Proceedings of international joint conference on artificial intelligence* (pp. 992–997).
- Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: bagging, boosting, and variants. *Machine Learning*, 36, 105–139.
- Bezdek, J. (1981). *Pattern recognition with fuzzy objective function algorithms*. New York: Plenum Press.
- Bezdek, J., & Hathaway, R. (1988). Recent convergence results for the fuzzy c-means clustering algorithms. *Journal of Classification*, 5(2), 237–247.
- Bingham, E., & Mannila, H. (2001). Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of international conference on knowledge discovery and data mining* (pp. 245–250).
- Boongoen, T., Shen, Q., & Price, C. (2010). Disclosing false identity through hybrid link analysis. *Artificial Intelligence and Law*, 18(1), 77–102.
- Boulis, C., & Ostendorf, M. (2004). Combining multiple clustering systems. In *Proceedings of European conference on principles and practice of knowledge discovery in databases* (pp. 63–74).
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24, 123–140.
- Celton, M., Malpertuy, A., Lelandais, G., & de Brevern, A. G. (2010). Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. *BMC Genomics*, 11, 15.
- Costa, J. A. F., & de Andrade Netto, M. (1999). Cluster analysis using self-organising maps and image processing techniques. In *Proceedings of IEEE international conference on systems, man, and cybernetics* (Vol. 5, pp. 367–372).
- de Castro L. N. (2001). *Immune engineering: development of computational tools inspired by the artificial immune systems*. Ph.D. thesis, DCA—FEEC/UNICAMP, Campinas/SP, Brazil, Department of Computer Engineering and Industrial Automation, School of Electrical and Computer Engineering, State University of Campinas, Brazil.

- Domeniconi, C., & Al-Razgan, M. (2009). Weighted cluster ensembles: methods and analysis. *ACM Transactions on Knowledge Discovery from Data*, 2(4), 1–40.
- Domeniconi, C., Gunopulos, D., Yan, B., Al-Razgan, M., & Papadopoulos, D. (2007). Locally adaptive metrics for clustering high-dimensional data. *Data Mining and Knowledge Discovery*, 14(1), 63–97.
- Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification* (2nd ed.). New York: Wiley-Interscience.
- Dudoit, S., & Fridyand, J. (2003). Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, 19(9), 1090–1099.
- Dunn, J. C. (1974). Well separated clusters and optimal fuzzy partitions. *Cybernetics and Systems*, 4(1), 95–104.
- Fern, X. Z., & Brodley, C. E. (2003). Random projection for high dimensional data clustering: a cluster ensemble approach. In *Proceedings of international conference on machine learning* (pp. 186–193).
- Fern, X. Z., & Brodley, C. E. (2004). Solving cluster ensemble problems by bipartite graph partitioning. In *Proceedings of international conference on machine learning* (pp. 36–43).
- Fern, X. Z., & Lin, W. (2008). Cluster ensemble selection. *Statistical Analysis and Data Mining*, 1(3), 128–141.
- Fischer, B., & Buhmann, J. M. (2003). Bagging for path-based clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(11), 1411–1415.
- Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2, 139–172.
- Fouss, F., Pirotte, A., Renders, J. M., & Saerens, M. (2007). Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 19(3), 355–369.
- Fred, A. L. N., & Jain, A. K. (2002). Data clustering using evidence accumulation. In *Proceedings of international conference on pattern recognition* (pp. 276–280).
- Fred, A. L. N., & Jain, A. K. (2003). Robust data clustering. In *Proceedings of IEEE international conference on computer vision and pattern recognition* (pp. 128–136).
- Fred, A. L. N., & Jain, A. K. (2005). Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6), 835–850.
- Fred, A. L. N., & Jain, A. K. (2006). Learning pairwise similarity for data clustering. In *Proceedings of international conference on pattern recognition* (pp. 925–928).
- Getoor, L., & Diehl, C. P. (2005). Link mining: a survey. *ACM SIGKDD Explorations Newsletter*, 7(2), 3–12.
- Ghosh, J., & Acharya, A. (2011). Cluster ensembles. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(4), 305–315.
- Gionis, A., Mannila, H., & Tsaparas, P. (2007). Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 4-ex.
- Hadjitodorov, S. T., Kuncheva, L. I., & Todorova, L. P. (2006). Moderate diversity for better cluster ensembles. *Information Fusion*, 7(3), 264–275.
- Han, J., & Kamber, M. (2000). *Data mining: concepts and techniques* (1st ed.). San Mateo: Morgan Kaufmann.
- Hu, X., & Yoo, I. (2004). Cluster ensemble and its applications in gene expression analysis. In *Proceedings of Asia-Pacific bioinformatics conference* (pp. 297–302).
- Iam-On, N., Boongoen, T., & Garrett, S. (2008). Refining pairwise similarity matrix for cluster ensemble problem with cluster relations. In *Proceedings of eleventh international conference on discovery science* (pp. 222–233).
- Iam-On, N., Boongoen, T., & Garrett, S. (2010). LCE: a link-based cluster ensemble method for improved gene expression data analysis. *Bioinformatics*, 26(12), 1513–1519.
- Iam-On, N., Boongoen, T., Garrett, S., & Price, C. (2011). A link-based approach to the cluster ensemble problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12), 2396–2409.
- Iam-On, N., Boongoen, T., Garrett, S., & Price, C. (2012). A link-based approach to cluster ensemble approach for categorical data clustering. *IEEE Transactions on Knowledge and Data Engineering*, 24(3), 413–425.
- Jain, A. K., & Dubes, R. C. (1998). *Algorithms for clustering*. Englewood Cliffs: Prentice-Hall.
- Jain, A. K., Duin, R., & Mao, J. (2000). Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1), 4–37.
- Jeh, G., & Widom, J. (2002). SimRank: a measure of structural-context similarity. In *Proceedings of international conference on knowledge discovery and data mining* (pp. 538–543).
- Karypis, G., & Kumar, V. (1998). Multilevel k-way partitioning scheme for irregular graphs. *Journal of Parallel and Distributed Computing*, 48(1), 96–129.
- Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: an introduction to cluster analysis*. New York: Wiley.

- Kim, D. W., Lee, K. Y., & Lee, K. H. (2007). Towards clustering of incomplete microarray data without the use of imputation. *Bioinformatics*, 23(1), 107–113.
- Kim, E., Kim, S., Ashlock, D., & Nam, D. (2009). MULTI-K: accurate classification of microarray subtypes using ensemble k-means clustering. *BMC Bioinformatics*, 10, 260.
- Kim, K., & Ahn, H. (2008). A recommender system using GA K-means clustering in an online shopping market. *Expert Systems with Applications* 34, 1200–1209.
- Kim, S., & Lee, J. (2007). Ensemble clustering method based on the resampling similarity measure for gene expression data. In *Statistical methods in medical research* (pp. 1–26).
- Kittler, J., Hatef, M., Duin, R., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3), 226–239.
- Kuncheva, L. I. (2006). Experimental comparison of cluster ensemble methods. In *Proceedings of international conference on fusion* (pp. 105–115).
- Kuncheva, L. I., & Hadjitodorov, S. T. (2004). Using diversity in cluster ensembles. In *Proceedings of the IEEE international conference on systems, man & cybernetics* (pp. 1214–1219).
- Kuncheva, L. I., & Vetrov, D. (2006). Evaluation of stability of k-means cluster ensembles with respect to random initialization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11), 1798–1808.
- Lam, L., & Suen, C. Y. (1997). Application of majority voting to pattern recognition: an analysis of its behavior and performance. *IEEE Transactions on Systems, Man and Cybernetics*, 22(5), 553–568.
- Law, M., Topchy, A., & Jain, A. K. (2004). Multiobjective data clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 424–430).
- Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7), 1019–1031.
- Lin, Z., King, I., & Lyu, M. R. (2006). Pagesim: a novel link-based similarity measure for the world wide web. In *Proceedings of IEEE/WIC/ACM international conference on web intelligence* (pp. 687–693).
- Mangasarian, O. L., & Wolberg, W. H. (1990). Cancer diagnosis via linear programming. *SIAM News*, 23(5), 1–18.
- McQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (pp. 281–297).
- Minaei-Bidgoli, B., Topchy, A., & Punch, W. (2004). A comparison of resampling methods for clustering ensembles. In *Proceedings of the international conference on machine learning: models, technologies and applications* (pp. 939–945).
- Minkov, E., Cohen, W. W., & Ng, A. Y. (2006). Contextual search and name disambiguation in email using graphs. In *Proceedings of international ACM SIGIR conference on research and development in information retrieval* (pp. 27–34).
- Mirkin, B. (2001). Reinterpreting the category utility function. *Machine Learning*, 45, 219–228.
- Monti, S., Tamayo, P., Mesirov, J. P., & Golub, T. R. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52(1–2), 91–118.
- Naldi, M. C., Carvalho, A. C., & Campello, R. B. (2013). Cluster ensemble selection based on relative validity indexes. *Data Mining and Knowledge Discovery*. doi:10.1007/s10618-012-0290-x.
- Ng, A., Jordan, M., & Weiss, Y. (2001). On spectral clustering: analysis and an algorithm. *Advances in Neural Information Processing Systems*, 14, 849–856.
- Nguyen, N., & Caruana, R. (2007). Consensus clusterings. In *Proceedings of IEEE international conference on data mining* (pp. 607–612).
- Punera, K., & Ghosh, J. (2007). Soft cluster ensembles. In *Advances in fuzzy clustering and its applications* (pp. 69–90). New York: Wiley.
- Reuther, P., & Walter, B. (2006). Survey on test collections and techniques for personal name matching. *International Journal on Metadata, Semantics and Ontologies*, 1(2), 89–99.
- Shi, T., & Horvath, S. (2006). Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*, 15(1), 118–138.
- Shi, T., Seligson, D., Belldgrun, A. S., Palotie, A., & Horvath, S. (2005). Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. *Modern Pathology*, 18(4), 547–557.
- Strehl, A., & Ghosh, J. (2002). Cluster ensembles: a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3, 583–617.
- Swift, S., Tucker, A., Vinciotti, V., Martin, N., Orengo, C., Liu, X., & Kellam, P. (2004). Consensus clustering and functional interpretation of gene-expression data. *Genome Biology*, 5, R94.
- Tan, P. N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Reading: Addison Wesley.
- Tijms, H. (2004). *Understanding probability: chance rules in everyday life*. Cambridge: Cambridge University Press.

- Topchy, A. P., Jain, A. K., & Punch, W. F. (2003). Combining multiple weak clusterings. In *Proceedings of IEEE international conference on data mining* (pp. 331–338).
- Topchy, A. P., Jain, A. K., & Punch, W. F. (2004a). A mixture model for clustering ensembles. In *Proceedings of SIAM international conference on data mining* (pp. 379–390).
- Topchy, A. P., Law, M. H. C., Jain, A. K., & Fred, A. L. (2004b). Analysis of consensus partition in cluster ensemble. In *Proceedings of IEEE international conference on data mining* (pp. 225–232).
- Topchy, A. P., Jain, A. K., & Punch, W. F. (2005). Clustering ensembles: models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12), 1866–1881.
- Wu, R. C., Chen, R. S., Chang, C. C., & Chen, J. Y. (2005). Data mining application in customer relationship management of credit card business. In *Proceedings of international conference on computer software and applications* (pp. 39–40).
- Xue, H., Chen, S., & Yang, Q. (2009). Discriminatively regularized least-squares classification. *Pattern Recognition*, 42(1), 93–104.
- Yu, Z., Wong, H. S., & Wang, H. (2007). Graph-based consensus clustering for class discovery from gene expression data. *Bioinformatics*, 23(21), 2888–2896.